MASTER

Predictive modeling using external data to predict the next best call

van den Broek, S.

Award date:
2014

Link to publication
Predictive Modeling in the Sales Funnel

Predictive Modeling Using External Data to Predict the Next Best Call

Van Den Broek, S.
9/23/2014

First supervisor: Remco Dijkman
Second supervisor: Rui Jorge Almeida
Supervisor I-bridge: Martin Dijkink
Abstract
This research presents an analysis of the effects of building and implementing a predictive model in the sales process measured in terms of the key performance indicator ‘call conversion’. Call conversion entails the calls made for outbound acquisition of companies potentially interested in temporary personnel. For companies in highly competitive environments the need to utilize the limited resources available more and more effectively is a key requirement to remain competitive. The costs have to go down while the performance must remain the same or even go up. In order to achieve this goal companies are trying to unlock the untapped value resting inside the vast amounts of data in their possession.
Nowadays, companies are slowly beginning to see that that potential value can be unlocked using predictive modeling. It often remains nothing more than a theoretical promise from data scientists that is rarely implemented in practice. This research will address this problem for one of the leading companies in the world wide staffing industry. A model has been built for estimating the probability of call conversion in the context of the sales process based on both internal and external data sources using a machine learning approach. This model has been field tested by using data from an actual call event. A call event is a sales campaign in which a selection of companies is called systematically with the goal of acquisition.
The research approach consisted of the following steps: First, the appropriate machine learning algorithms to build the models have been collected from literature. Second, the relevant part of the sales process has been analyzed using an interview approach. These interviews yield a usable definition for call conversion and a set of potentially fruitful external data sources. Thirdly, data has been gathered and preprocessed to become ready for the machine learning algorithms. Due to very imbalanced data specific performance measures had to be collected from literature and constructed to assess the performance of the models properly. The data set is then used to build and optimize the model using two algorithms: the decision tree and the random forest. Finally, the model has been validated using data from an actual call event. The results of the best model compared to the random set of the validation are described in Table 1. The conversion of the random set on the first row is the call conversion on a randomly selected set of companies. This is compared to the conversion on a set of companies selected by a model. The best conversion that is achieved using a decision tree model. When these two values are compared it can be seen that a 43% increase in conversion is achieved.

<table>
<thead>
<tr>
<th></th>
<th>Conversion</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random set</td>
<td>7.9%</td>
<td>-</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>11.3%</td>
<td>43%</td>
</tr>
</tbody>
</table>

Table 1: The results of the validation using the data from the call event.

The model improves performance even though the set of companies used in Table 1 has more specific characteristics than the set of companies used to build the model. To be able to use the model on this data several assumptions had to be made about the used definition of conversion and the external data. The model still produces positive results which can be seen as an indication of robustness of the model. This research provides not only results that shows the promise of applying predictive modeling in the sales process but also shows the promise of the potential value of using this approach to unlock value.
hidden in the vast amounts of data that companies possess nowadays. These results can thus be seen as a starting point of a whole collection of predictive models that are used throughout different business processes. For practitioners it can be used as a justification of the need for a proper architecture that is able to support these types of models and actually implement them in the business processes. The potential value of using external data has also been tested in this thesis and it has shown signs that this might be useful. This research thus serves as an indication that predictive modeling is valuable in the context of the sales process of the staffing industry. Apart from that this research can serve as a starting point for further research in the usefulness of external data in predictive modeling. This can be done both in the sales process of other industries or in other processes within the staffing industry.
Preface
This research is a master thesis and serves as a conclusion of the master study Operations Management & Logistics at the Eindhoven University of Technology with the Information Systems Group at the department of Industrial Engineering & Innovation Sciences.

The project is done in cooperation with iBridge B.V. in Amsterdam which is the shared service center IT for Randstad Netherlands N.V. The research has been done within the Market Exploration team although input has been received from various other departments within Randstad. Randstad is one of the world’s leading staffing agencies that basically invented the concept of staffing. Apart from iBridge, the Information Systems Group at the faculty of Industrial Engineering & Innovation Sciences at the Eindhoven University of Technology has been involved through guidance, feedback, and support. The project has taken a total of 6 months.

My personal goal with this master thesis was to get more familiar with the emerging and even hyping field of Big Data. Especially the applications of the typical Big Data techniques in industries that are less IT-focused but do possess huge piles of data. In my opinion this poses the biggest opportunities for value creation if the organization is able to deal with the change in thought processes needed. The combination of business challenge along with a quantitative challenge proved to be a great learning experience. It gave me the opportunity to use both the skills I acquired during my studies as well as the skills I acquired during extracurricular activities.

The progress for the thesis has almost gone as planned out, since the planning was made in an iterative way. So plans were flexible in the sense that new plans were made whenever a hitch presented itself. The interviews have been performed on a convenience approach. It has been very motivating to experience the degree of readiness to help from people across the organization. This ranged from blocking time in their agendas to running queries.

For preprocessing the data and building the model different tools have been used. The initial try was to use Microsoft Excel but this normally quite versatile tool utterly fails when the amount of data grows even a little bit. Finally the preprocessing has been done in a tool called Alteryx of which several trial versions proved sufficient to fulfill its purpose. The model building has been done in the open source tool Knime in combination with the WEKA plugin.

Finally, I would like to thank the supervisors who helped me with this thesis: From Randstad I would like to thank my direct supervisors Martin Dijkink, Alexander Croiset, Peter Zijnstra, and Martijn Imrich. Apart from that I would like the great variety of people who have helped me throughout the process. From the Eindhoven University of Technology I would like to thank Remco Dijkman and Rui Jorge Almeida.

But my biggest appreciation of all goes to Ilse and my parents, who have always been there for me when I needed it and also just because, motivating me all the way, who made me proud to have them in my life, and without whom this thesis would not be what it is today.

Sander van den Broek
Table of Contents
Abstract................................................................................................................................. ii
Preface ................................................................................................................................. iv
List of Figures ....................................................................................................................... vii
List of Tables ......................................................................................................................... vii

1. Introduction ....................................................................................................................... 1
   1.1. Problem outline and relevance...................................................................................... 1
   1.2. Key concepts and definitions ....................................................................................... 2
       1.2.1. Business performance......................................................................................... 2
       1.2.2. Predictive analytics............................................................................................. 4
   1.3. Research goal and descriptions .................................................................................... 5
   1.4. Research design .......................................................................................................... 7

2. The different techniques and tools for predictive analytics ............................................ 8
   2.1. Prediction techniques ................................................................................................ 8
       2.1.1. Decision trees.................................................................................................... 8
       2.1.2. Random forest ................................................................................................ 9

3. Interviews to determine definition and data sources .................................................... 12
   3.1 Introduction to the in depth interview approach ......................................................... 12
   3.2 In depth interview framework .................................................................................... 13
   3.3 Interview questions .................................................................................................... 14
   3.4 Case selection ............................................................................................................. 15
   3.5 Research instruments ................................................................................................. 15
   3.6 Conduct and transcribe interviews .......................................................................... 16
   3.7 Data coding .............................................................................................................. 16
   3.8 Check data completeness and review data ............................................................... 18
   3.9 Analyze interview data ............................................................................................. 18
       3.9.1 Call conversion definition ................................................................................. 18
       3.9.2 sources of data.................................................................................................. 20

4. Data ................................................................................................................................ 22
   4.1 Exploration of the data .............................................................................................. 22
   4.2 Imbalanced data ......................................................................................................... 26
       4.2.1. Two techniques for assessing the imbalanced data problem.............................. 26
List of Figures

Figure 1; The performance management framework. (Ferreira, 2012) ................................................. 4
Figure 2; The conceptual process of implementing a predictive application (Gualtieri, M., 2013) .......... 5
Figure 3; The research design to structure the thesis ................................................................................. 7
Figure 4; An example of a decision tree ...................................................................................................... 9
Figure 5; The statistical problem explained (Polikar, 2008) ...................................................................... 9
Figure 6; Graphical illustration of the representational problem (Polikar, 2008) ...................................... 10
Figure 7; The overview of the interview approach ....................................................................................... 13
Figure 8; The procedure for selection of the interviewees ................................................................. 15
Figure 9; The accumulated percentage of days between call and request versus number of days and days between call and visit versus number of days ....................................................................................... 19
Figure 10; The pre-processing steps from raw data to pre-processed set ................................................ 22
Figure 11; The hypothetical test used for the validation ........................................................................... 28
Figure 12; An example of an ROC curve .................................................................................................. 28
Figure 13; The general flow built in Knime ............................................................................................... 31
Figure 14; Decision tree versus the random forest, max costs saved measurement with all the data used. ........................................................................................................................................ 33
Figure 15; Decision tree versus random forest, max f-value with all the data ........................................... 34
Figure 16; The ROC curves for max cost savings using only internal data ............................................. 35
List of Tables

Table 1; The results of the validation using the data from the call event ............................................. ii
Table 2; The interview questions ........................................................................................................... 14
Table 3; The coding of the interviews: the definition of call conversion Y= yes, N= no ...................... 16
Table 4; The coding of the interviews: the sources ................................................................................. 17
Table 5; An example of the used data. ....................................................................................................... 25
Table 6; The relevant metrics for assessing model performance (He & Garcia, 2009) ......................... 28
Table 7; The different parameters of the model building. ....................................................................... 30
Table 8; The results of the best models using all data.............................................................................. 33
Table 9; The results for the model using the internal data only............................................................... 35
Table 10; The results of the validation .................................................................................................... 37
Table 11; The best models per technique and per performance measure .............................................. 39
Table 12; The results of the validation .................................................................................................... 40
Table 13; The different models that have been built for the decision tree ........................................... 49
Table 14; The results of the decision tree ............................................................................................... 50
Table 15; The settings of all the models that have been tested for the random forest ....................... 52
Table 16; The results for the random forest models .............................................................................. 53
Table 17; The results of the earlier external data .................................................................................... 53
1. Introduction
This thesis looks at predicting call conversion in the sales funnel, using external data sources in addition to the internal data sources employing machine learning models. Several prediction algorithms are applied to build a model to enhance business performance in terms of call conversion. The research is done in the context of the staffing industry. Call conversion entails the calls that are made for outbound acquisition of companies potentially interested in temporary personnel.
In section 1.1 the research problem and the relevance of the research are discussed in some detail. In section 1.2 the important concepts are discussed and defined. In the subsequent section 1.3 the research problem is revisited and is detailed in the form of research questions and the splitting the research goal into several sub-goals. Finally, in section 1.4 it is described how the research goals will be achieved by means of providing a research design which will be used throughout the thesis. On the basis of section 1.4 the rest of the report is based.

1.1. Problem outline and relevance
In this section a preliminary description of the research field is given. This is followed by a problem statement and description of further steps.
In highly competitive environments companies need to focus more and more on utilizing the resources available effectively to remain competitive. The costs have to go down while the business performance must remain the same of even increase. In order to do this different processes have to be continuously monitored and improved. This has traditionally been done by trying to control the processes in a quantitative way. From this the approach of business intelligence originated. Business intelligence is defined as a set of theories, methodologies, architectures, and technologies that transform raw data into meaningful and useful information for business purposes (Rud, 2009). In business intelligence dashboards are often used to display the performance. These dashboards in general show measures from the past and can be considered big rear view mirrors. The next step in performance management is looking forward into the future with prediction algorithms (TDWI, 2013). The market of predictive analytics tools is maturing, generating more powerful tools that are easier to use (TDWI, 2013). The combination of these two factors clearly shows the relevance and potential of the use of predictive analytics in both this research as well as in practice. Predictive analytics in the context of acquiring new customers has been researched before. For instance, it has been shown that it is possible to predict profitability of a suspected customer in a business-to-business context (D'Haen, 2013). In that same report, it is also indicated that this can be augmented by predicting which customer to attract on the basis of the chance that it converts. To the best knowledge of the author, this specific approach applied in this industry has not been extensively documented in research yet. In the context of the staffing industry the value of predictive modeling can be particularly big since the service that is being sold is complex in nature, i.e. it is hard to tell which company needs which employee at which time. Also the demand from the market is strongly tied to the shape of the economy. The combination of these factors make it very hard to manually determine the right company to call from the huge amount of potential customers. So, in the everlasting quest to remain competitive and acquiring enough customers, the company has decided to look into the application of predictive analytics. The combination of these factors make it an interesting gap in research and a promising direction in practice to look into.
Several years ago, advanced predictive analytics was a technique applied predominantly by statisticians and researchers. With the development of more off the shelf and easy to use predictive analytics tools comes the movement to democratize and unlock predictive analytics for the whole business. However, it is still common practice that the data scientists build the model and handle the data, that the business use the model, and that the final users only see the results without knowing that there is a complex model behind it. This last part remains key, since it is very hard to change behavior of people when it comes to new solutions and products (Gourville, 2006). On top of this, it is also found from practice that these predictive solutions used to be following technology push and recently this is changing to a business pull. Finally, the availability and accessibility of data from different sources is growing in a spectacular fashion. It is becoming feasible to look beyond the borders of the company database and incorporating external data sources.

When taken the previous developments into account, the following problem statement can be formulated.

### Problem statement

*It is unclear what the effect of using external data in addition to the traditional internal data on sales prediction is. Furthermore it is unclear how predictive analytics in sales perform in enhancing business performance in practice.*

This problem will be analyzed in the context of the staffing industry. This is an interesting industry since the product they sell is very complex and very dependent of the state of the economy. The focus will be on the business performance in the business process of sales, i.e. the sales funnel.

First, the necessary techniques will be described. After that, an exact definition of call conversion will be derived and suitable external data sources will be chosen on the basis of consulting people from practice. Combining these two, a predictive model will be built and optimized. Finally, this model will be field tested in the business process.

The main focus of this thesis lies in building a model that can predict the call conversion using external data sources on top of the traditional internal data sources. The internal data sources contain the data about the execution of the business processes. Another challenge is to show how this predictive model can be used to improve business performance.

### 1.2. Key concepts and definitions

To start addressing the problem properly, first important definitions have to be described. First, the broader concept of business performance and how sales conversion fits in this concept is described. Second, the concept of predictive analytics is elaborated upon in this section.

#### 1.2.1. Business performance

In the modern complex and dynamic business environment companies need to be able to adapt to increasingly rapid market changes (Ferreira, 2012). This leads to the tendency of companies to focus on their core competences and the factors that really determine that competitive advantage. These developments lead to an increased importance of business networks that function well and meet the
objective of meeting customers’ needs efficiently and effectively (Chituc & Azevedo, 2005). In order to enable these networks to function properly, a set of measures that gives a complete and correct view of the current business in terms of financial and non-financial measures, internal and external measures, and efficiency and effectiveness measures is needed (Kaplan, 1996). This is done using performance measurement.

Performance measurement has been defined as consisting of the three following interrelated elements (Neely, 1995).

- Individual measures that quantify the efficiency and effectiveness of actions.
- A set of measures that combine to assess the performance of an organization as a whole.
- A supporting infrastructure that enables data to be acquired, collated, sorted, analyzed, interpreted, and disseminated. This performance measurement is usually done by collecting data and representing them using Key Performance Indicators (KPI’s) and Critical Success Factors (CSF’s).

The definition of key performance indicator is as follows:
KPI’s are quantifiable metrics which reflect the performance of an organization in achieving its goals and objectives (Bauer, 2004).

A ‘critical success factor’ is defined as follows (Bauer, 2004):
Critical success factors are those few things that must go well to ensure success for a manager or an organization, and, therefore, they represent those managerial or enterprise area, that must be given special and continual attention to bring about high performance. CSFs include issues vital to an organization’s current operating activities and to its future success (Boynton & R.W., 1984).

The biggest difference between a KPI and a CSF is that a KPI is quantitative in nature and that a CSF is qualitative in nature. KPI’s are often an indirect measurement of the CSF’s. In the context of this thesis the KPI call conversion will be addressed. Call conversion is the conversion on outbound calls that are intended to get companies to place orders for temporary staff. An exact definition will be stated later on in the project when more data exploration has been done.

A framework describing the relation between various performance indicators and their stakeholders has been derived (Ferreira, 2012) and can be seen in Figure 1.
This framework consists of three levels. The first and most important level is to identify the strategy and vision of the network, which is followed by the executing and monitoring, and output analysis. The second level is monitoring and executing. This means that the relevant performance measures are discussed and defined in order to enable effective monitoring. This also entails the defining of target values and subsequently optimizing processes. The third and final level is output analysis where the performance of the business network is analyzed and then it is tried to improve the performance of the whole business network. This thesis takes place in the output analysis and execution & monitoring phase.

1.2.2. Predictive analytics
Predictive analytics has been gaining in popularity in the past years. It is defined as a variety of statistical techniques from modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future, or otherwise unknown, events (Nyce, 2007).
For this thesis, the focus will be on two types of machine learning models: random forests and decision trees. This focus has been chosen because these models provide strong performance and are used often in similar studies (D'Haen, 2013).
What will be described here is the conceptual process that needs to take place when building and implementing a predictive application, as can be seen in Figure 2.
The process is a continuous process which means that firms need to rerun the cycle constantly in order to maximize business value from the predictive applications. The process has similarities with the structure of this thesis. In this thesis one sequence of this cycle and small iterations between elements will be attempted.

1.3. Research goal and descriptions
On the basis of the research problem a research goal is derived which in turn will be divided into several sub goals. These goals will be translated into several research questions. When these questions are answered separately this should aid in achieving the overall goal.

<table>
<thead>
<tr>
<th>Overall goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Build a model to predict the call conversion using external data sources on top of the traditional internal data sources. Secondly, implement this prediction model to enhance business performance.</td>
</tr>
</tbody>
</table>

This overall goal falls apart into several sub goals. First, it is important to find external data sources that are useful for the prediction model. Secondly, it is key to build and compare different predictive models that predict call conversion. Following the results of the first two goals, the goal is to implement the resulting model and validate its effect.
These three sub-goals are mapped to three research questions.
Research question 1:
What data sources are suitable in the context of predicting the call conversion?
1a: What external data sources can be used to predict the value of the call conversion?
1b: What is a usable definition for call conversion and how is this ‘datafied’ internally?

Research question 2:
What is the best predictive model that can be built to predict call conversion using the results of RQ1?

Research question 3:
What is the effect of using the predictive model in practice on business performance in terms of call conversion?
1.4. Research design

In the previous paragraphs the goals and sub-goals have been defined. On the basis of these goals a research design has been made, which will provide a framework to guide the answering of the research questions. The research design is illustrated in Figure 3.

Figure 3; The research design to structure the thesis.

The first research question is about gaining insight into the business context and consists of two parts. This will be answered in a qualitative fashion by gathering information from practitioners by means of conducting interviews. On the basis of the process and data knowledge of these expert practitioners, potential relevant data sources are selected. The description can be found in chapter 3.

The second question will be about building a predictive model incorporating the techniques found in literature and the data sources and the call conversion and will thus be answered in a quantitative fashion. The literature has already been described in chapter 2. The rest will be described in chapter 4, chapter 5, and chapter 6.

The third and final question is about implementing the model into practice and evaluating the results and will again be answered in a quantitative fashion. This will be described in chapter 7.
2. The different techniques and tools for predictive analytics

In this chapter the different techniques and tools for predictive analytics as used in the context of this thesis are described. The results of the models will be compared to the current situation in which no model at all is used. Because of this obtaining the best performance theoretically possible is not of utmost importance. For this thesis, the focus will be on two types of models: decision trees and random forest. This focus has been chosen because of several reasons. First, these models are used in similar studies (D’Haen, 2013; Chang, 2011). These algorithms have been compared and both show strong performance and are easily interpretable (Khan, Ali, Ahmad, & Maqsood, 2012). From both of the algorithms instances are available that are able to handle many different types of data so the pre-processing effort will be feasible (Khan, Ali, Ahmad, & Maqsood, 2012).

2.1. Prediction techniques

For this thesis, two types of classification algorithms are used.

- Decision trees
- Random forest

The algorithms are sufficient for this study and are all implementable into one of the available data mining tools. Throughout the years a vast amount of research on prediction techniques in general and these algorithms specifically has been published. In this paper the focus is on applying the algorithms, so the description will not go beyond introductory depth. The exact mathematical description will not be given here. Sources where this can be found will be referenced in the following sections. This suffices because most complexity can be handled by the tools used to build the models.

First, the decision tree model is described and secondly the random forest ensemble method is described.

2.1.1. Decision trees

A commonly used method for prediction models is decision tree learning (Quinlan R. J., Induction of decision trees, 1986). It is based on the structure of a tree: the leaves represent class labels and the branches represent the conjunctions the paths that lead to those class labels. The main advantage is that it is a clear way to provide a visual depiction of the model, thus making the dependencies easy to understand and interpret.

When looking at Figure 4, an example of a decision tree is given. Each of the internal nodes represents one of the input variables.

The concept of a decision tree consists of different steps. At each node a choice for the best attribute to explain the split is made. The way ‘best’ is determined depends on the algorithm that is used. Often this step is done by some sort of information theory test (Han, Kamber, & Pei, 2012). Most algorithms follow a greedy top-down recursive divide-and-conquer manner for constructing decision trees. This means that a top-down induction of variables is used.

Many different algorithms for constructing decision trees are available like C4.5 (Quinlan R. J., 1993), CART (Breiman, Friedman, Olshen, & Stone, 1984), CHAID (Kass, 1980), and ID3 (Quinlan R. J., 1986). The
The difference between these models is the derivation of the specific rules of the model that result in splitting, like entropy and information gain. In the context of this thesis, C4.5 will be used because this is often used in direct marketing (Ratner, 2012) and performs well (Quinlan R. J., 1996). This method makes an analysis on how the variables best combine to explain the outcome of the dataset. This is done on the basis of a statistical test (Han, Kamber, & Pei, 2012). One main advantage of C4.5 is that it is non-parametric and hence it does not require that the underlying data is normally distributed. It also has the ability to handle a big amount of variables and missing values (Ratner, 2012).

![Decision Tree Example](image)

Figure 4: An example of a decision tree

Figure 4 is an example of a decision tree built using the c4.5 algorithm. It concerns buying decisions of stocks. The algorithm finds that stock levels have the highest information gain so it is put on top of the tree. If the stock level is below ten, then the number available has the highest information gain. If in that case, the number available is below ten, then the buy all prediction is made. The same way of reasoning goes for the other branches.

### 2.1.2. Random forest

In the above algorithm, that looks for one single hypothesis, typically three types of problems can occur: a statistical problem, the computational, and the representation problem. These three problems can be partly overcome by ensemble methods, methods in which multiple models are used in combination with each other (Dietterich, 1997). The statistical problem emerges when a learning algorithm has insufficient training data.

![Random Forest Diagram](image)

Figure 5: The statistical problem explained (Polikar, 2008)
to search a space of hypotheses. As a result an algorithm can find different 'optimal' solutions and consequently can choose the wrong one. The statistical problem is illustrated in Figure 5. Imagine a two dimensional data set which has to be categorized into three categories. The upper three figures depict three separate classifiers that each have non-perfect performance but are close. When these are combined using an ensemble method, which takes a combination of the three separate classifiers, it can be seen that the real decision boundary is approached and this increases the performance. Using a voting technique, like in many ensemble methods, this can be achieved.

The representational problem emerges when the hypotheses space does not contain a hypothesis that is a suitable approximation of the true function. With a weighted sum of these hypotheses it is sometimes possible to expand the hypotheses space. This enables the algorithm to find a better approximation of the true function. The representational problem is displayed in Figure 6. A two dimensional problem with two categories is displayed (circles and crosses). The red line is the true function that divides the two categories. This complex function cannot be described by one single classifier (the circles). When combining multiple of those classifiers, the complex boundary can be approached.

The computational problem arises when the learning algorithm cannot guarantee to find the best hypothesis within the hypotheses space. For instance, decision trees are impossible to train perfectly so heuristic methods are used. Because of this the algorithm can get stuck in local minima and hence fail to find the best hypothesis. Employing a weighted combination of multiple different classifiers can reduce this risk.

Commonly used ensemble methods manipulate the training sets and run the learning algorithm several times. This works especially well for unstable algorithms like decision trees (Dietterich, 2000). Examples of this type of ensemble methods are bagging, AdaBoost, and Random Forests. The term bagging is derived from ‘bootstrap aggregation’ and it is the most straight-forward method. It feeds the learning algorithm with a training set consisting of a sample drawn with replacement from the original training set. Such a set is called a bootstrap replicate. The AdaBoost algorithm assigns weights to the training examples. The learning algorithm is trained with this altered training set and thus minimizes the weighted error on the training set. This procedure is repeated and basically more weight is placed on training examples that were classified incorrectly and less weight on training examples that were classified correctly. This ultimately minimizes the error function (Dietterich, Ensemble Methods in Machine Learning, 2000).

The AdaBoost algorithm is basically a deterministic algorithm as it grows the trees by successive reweighting of the training set. Because of this there is a relatively strong correlation between different
member trees within the algorithm. This is hurting the accuracy (Breiman, 2001). The Random Forest algorithm uses randomization instead of the successive reweighting (Breiman, 2001). Random Forest combines the method of bagging with the random selection of features to include. First the bagging technique is applied to tree learners which means that a random subsets of the training data are taken and used to generate one tree per sample. After that a random set of attributes is selected at each split of each tree. This ensures that there is little correlation between the different member trees thus enabling high generalizability. Therefore, the Random Forest algorithm will be used in the context of this thesis.
3. Interviews to determine definition and data sources
The goal of this part of the research is to add contextual information to the literature described earlier. The interviews result in an exact definition of the call conversion and a list of potential external resources. These results help guide and focus the further research. This part will answer the second research question.

The interviews are done using a structured approach based on the in depth interview approach of (Boyce & Palena, 2006). This approach is used because it is useful when context of new issues is being explored or when a more complete picture of a phenomenon is required (Boyce & Palena, 2006).

First, in section 3.1 this approach is further explained. Secondly, in section 3.2 to 3.5 the data collection preparation for the interviews is described. Section 3.2 goes into the context of the in depth interviews and the research framework, section 3.3 goes into the interview questions, 3.4 into the case selection, and 3.5 into the research instruments used. The next stage, i.e. data collection and presenting the findings are discussed in section 3.6 to 3.8. Section 3.6 will focus on doing the interviews and transcribing them, section 3.7 will focus on checking data completeness and reviewing of the interviews and 3.8 focusses on coding the interviews. In the final part of this chapter the analysis, the definition of call conversion and the different data sources are formulated in section 3.9 and 3.10 respectively.

3.1 Introduction to the in depth interview approach
The in depth interview approach has been described elaborately in literature (Boyce & Palena, 2006; Guion, Diehl, & McDonald, 2011). Their methodologies are not identical but overlap and complement each other. Each of them focuses on different parts of a general action plan. The applicable part of the in depth research action plan can be split into three main steps.

- Data collection preparation
- Data collection, checking and presenting
- Analysis

When the research framework and the context have been described the questions for the interviews can be determined. After that, a selection of people from the business that should be approached will be made. The final step of the preparation phase is the selection of the research instruments.

For the next phase, the data collection, checking, and presenting step, the actual data is collected by conducting and transcribing the interviews. The textual data will be coded so that it is prepared to detect relations and variables. After the coding, the data will be checked with the interviewees to ensure completeness and correctness. This result will be presented in a table.

With the coded data the final step can be executed: the analysis. The result of the analysis will be a definition for call conversion and an overview of the useful sources. In Figure 7 the overview of the method is summarized.
3.2 In depth interview framework

Four different people from different parts of the organization have been interviewed. These people work and interact with the call conversion on all different kinds of levels across the organization, reaching from operational personnel up to business intelligence managers from different business units. In this way a complete view from different perspectives on call conversion is guaranteed. This complete view is necessary to limit bias, being one of the biggest pitfalls from the in depth approach (Boyce & Palena, 2006). From the perspective of the IT organization of Randstad group NL, i-Bridge, the operational units and the different business units are customers. This set of people has been selected because a variety of practical reasons:

- Direct access to these people is available
- They interact with the subject from different perspectives
- The view of the customers can be compared with the view of i-Bridge
For this interview framework the position on in depth interviews is summarized as the ‘traveler metaphor’ (Legard, Keegan, & Ward, 2003). This metaphor falls within the constructivist research model, i.e. knowledge is created through the interaction between the interviewer and the interviewee. This is applicable to this study since no explicit usable definition for call conversion exists inside the organization so this has to be created from scratch.

3.3 Interview questions

The formulated questions can be found in Table 2. Although the interviews were conducted in Dutch the questions are listed in English. The questions are derived with three goals in mind:

- Context elaboration
- Elaboration on the definition of call conversion
- Finding (external) data sources that might be relevant.

The first category is about gaining insight into the context of the organization. The nested approach is used within one organization, that consists of several separate divisions. So it is necessary elaborate on the context.

The second category of questions aims to get a solid definition of call conversion. As a basis of these question the earlier experience in the organization gained for this thesis has been used. The last questions are about finding out which external data sources can be used to predict call conversion. One potential pitfall of this question is that the interviewees are very much focused internally and have a hard time thinking out of the box. In order to tackle this a question that tries to open up their minds has been added.

The information gaps that remain after the interviews have been transcribed and coded will be filled by emailing the interviewees with the request for the extra information.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Context elaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>What is your role inside the Randstad N.V. organization?</td>
</tr>
<tr>
<td>2</td>
<td>What is your involvement in acquiring new customers?</td>
</tr>
<tr>
<td>3</td>
<td>How does the sales funnel for acquiring new companies look?</td>
</tr>
<tr>
<td>4</td>
<td>How would you place call conversion into the sales funnel?</td>
</tr>
<tr>
<td>5</td>
<td>What calls would you take into consideration for call conversion?</td>
</tr>
<tr>
<td>6</td>
<td>When would you consider a call a success?</td>
</tr>
<tr>
<td>7</td>
<td>If no limitations existed on what you could measure, what would you want to know to predict call conversion?</td>
</tr>
<tr>
<td>8</td>
<td>What internal sources do you think predict call conversion?</td>
</tr>
<tr>
<td>9</td>
<td>How would you categorize the external data for predicting call conversion?</td>
</tr>
<tr>
<td>10</td>
<td>Do you use predictive models at the moment?</td>
</tr>
</tbody>
</table>

Table 2: The interview questions
3.4 Case selection
The selection of cases is mainly based on the availability of the relevant people for a face to face meeting. The selection process is done in two parallel steps as depicted in Figure 8. As an initial step direct colleagues are asked what people might be interesting to ask what might be relevant people to talk to (step A). This is done because then an introduction can be made, which enhances the chance that somebody is willing to make time for an interview. Secondly, people that have actually been interviewed are asked what other people might be interesting.

![Diagram of case selection process]

3.5 Research instruments
For gathering more information from the experts, interviews have been chosen as the method for data gathering. An interview takes approximately two hours.

All the interviews were done in Dutch and in a face-to-face setting. All the interviews were recorded and later transcribed to text. The latter yields a literal description of what has been said and is used for analysis purposes and for increasing the reliability of the research.
3.6 Conduct and transcribe interviews

The interviews were conducted over the course of several weeks with four interviews taking place being equally spread over this period of time. The interviews were transcribed as quickly as possible after ending the interviews, these transcripts can be found in Appendix II. During the transcription process no information gaps were discovered. However, after each interview a better understanding of the business was gained. This better knowledge was tested by asking earlier interviewees about their opinion about the development.

3.7 Data coding

The goal of the coding of the data is to transform the information of the transcripts of the interviews into a table with quantitative data. Based on the goal of the interviews, the coding is split into two parts plus context. The first part is the context of the interviewee, the second part is about the definition of call conversion, and the last part is about the external data sources. Both the finding a definition and the finding of the sources are started with as little preconceived vision about what the answer should be. Generally speaking, one would say that call conversion is the percentage of outbound calls, intended to generate revenue, that have positive follow-up thus taking the targeted company further into the sales funnel. However, this is not a feasible definition. For determining what ‘further in the sales funnel’ is the business-2-business side of the sales funnel used in the company is taken into account. This sales funnel is shown in Figure 9. The different call conversions that are considered are in the figures depicted by different arrows. The reason these are taken into account is since the other possible conversion are not stored consistently in the systems. The call conversions are call to recall, call to visit, call to request, and call to placement.

The steps to create a usable definition using the interview data are depicted in the Table 3 and Table 4.

![The B2B side of the sales funnel of Randstad](image)

**Table 3: The coding of the interviews: the definition of call conversion**

<table>
<thead>
<tr>
<th>Function</th>
<th>Marketing database manager</th>
<th>Business consultant</th>
<th>Manager Marketing Intelligence</th>
<th>Market intelligence analyst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company</td>
<td>Randstad NL</td>
<td>i-Bridge, a Randstad company</td>
<td>Tempo-Team, a Randstad company</td>
<td>Randstad NL</td>
</tr>
<tr>
<td>Definition</td>
<td>call to recall</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>call to visit</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>call to request</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>call to placement</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Use market units</td>
<td>Y</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Use account unit</td>
<td>Y</td>
<td>N</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Use xxl unit</td>
<td>N</td>
<td>N</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 9; The B2B side of the sales funnel of Randstad
In Table 3 the coding of the information relevant for the definition of call conversion is displayed. There are four rows that are about what to consider conversion. The final three rows are about which units to take into account. Each of these units serve a different kind of customer. The market unit serve small customers with a wide variety of requests. The account units serve the bigger customers, that often have a formal contract about the terms of delivery. The XXL units serve bigger customers with just a couple of different types of employees that are required very often. From the different interviews a structure of the definition became clear. This structure consists of the elements in the second column of Table 3. Each of the experts has been asked about their opinion about the inclusion of these elements into the data structure. The answers can be ‘yes, this should be included’ (Y) or ‘no, this should not be included (N).

<table>
<thead>
<tr>
<th></th>
<th>Marketing database manager Randstad NL</th>
<th>Business consultant i-Bridge</th>
<th>Market Intelligence Analyst Randstad NL</th>
<th>Manager Marketing Intelligence Tempo-Team</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Industry</td>
<td>Location</td>
<td>Time</td>
<td>Industry</td>
</tr>
<tr>
<td>BVNL¹</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Jobfeed²</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Side-inflow</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Nielsen³</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ROA³</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>VAT-filings of all companies</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Purchasing’s managers index⁴</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Temp staffing per industry</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>No. of listings per SBI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Business cycle survey⁵</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: The coding of the interviews: the sources

In Table 4 the coding of the different sources that might be useful for the analysis is shown. The structure is as follows: each of the sources named by the interviewees is listed on the left. On the rows it is shown on which parts of the internal data these sources can be mapped. For example, the source ‘Nielsen’ can be connected to the calls by using the industry or the time according to the Marketing database manager of Randstad NL. For each of the data sources the possible mappings as pointed out by each of the interviewees has been described.

1 www.cendris.nl
2 www.jobfeed.nl
3 www.nielsen.com
4 www.roa-maastricht.nl
5 https://www.nevi.nl/dossiers/pmi
6 www.cbs.nl
3.8 Check data completeness and review data
To check if the data was complete the coding of the interviews are sent to the interviewees. Note that the coding is sent instead of the transcripts as the latter would place another burden on the schedule of the interviewees. Several questions and requests for clarification were sent after the coding of the interviews. These results have been included in Table 3 and Table 4.

3.9 Analyze interview data
The analysis of the results of is split into two parts. First, the definition of call conversion will be given. Secondly, the different data sources that will be used for building the model will be described. These two parts are based on the results of the interviews, Table 3 and Table 4 respectively.

3.9.1 Call conversion definition
Since the concept of call conversion is one of the corner stones on which the whole thesis is built a separate section has been devoted to it. The practical implementation of this definition is described. The definitions are based on the interviews that have been described earlier. The practical implementation of the definition is based on explorations of the available data in combination with the coding of the interviews.
In practice not all information can be derived because of lacking data and imperfect data. The definition is given below.

‘Call conversion is defined in practice as the percentage of outbound calls to targeted companies (either suspective, prospective or inactive companies), intended for acquisition, that takes a targeted company further into the sales funnel by either a visit or a staffing request within 28 days.’

Note how the scoping of the definition only changed the wording and not the contents.
First, the concept of targeted companies is split into four groups. These four groups have been taken from the data directly. Three of these four groups are included in the definition. The active customers are not included in the definition since visiting an existing customer does not necessarily imply success. Including them would introduce a bias since it is expected that existing customers have a higher chance to stay customer than non-customers becoming customers.

- Suspective companies: companies with whom transactions have taken place more than 52 weeks ago.
- Prospective companies: companies with whom transactions have taken place more than 26 weeks ago but less than 52 weeks ago.
- Inactive companies: companies with whom transactions have taken place more than four weeks ago but less than 26 weeks.
- Active companies: companies with whom transactions have taken place in the past 4 weeks.

Calls get logged into the system by the sales representatives, or the interagents, that make the calls. Calls that are of interest for call conversion get logged under acquisition. From the interviews it became clear that not all relevant calls get logged under this header. However, if the scope gets broadened beyond acquisition the opposite happens and irrelevant calls come into scope. The latter is deemed to
be worse for the validity of the definition of call conversion. Underlying this is the assumption that whether or not an interagent logs calls properly does not influence call conversion of that interagent. Therefore the calls used are narrowed down to the category of acquisition, taking some false negatives for granted.

The next part of the definition is about taking the company further into the sales funnel as described in the earlier chapters. There are four conversions mentioned in the interviews as depicted in Table 3

- Call to recall
- Call to visit
- Call to request
- Call to placement

![Figure 10](image-url)  

Figure 10: The accumulated percentage of days between call and request versus number of days and days between call and visit versus number of days.

A conversion is included into the definition if more than half of the interviewees claims it is useful. When looking at Table 3, this means that at least three people indicate it is useful. This results in the inclusion of call to visit and call to request into the definition.

The next question that rises is how ‘call to visit’ and ‘call to request’ can be measured. This is not done
by using the coding of interviews as no conclusive answer was given. It is done by looking at the available internal data. A call is said to have resulted in a visit if a visit is registered within 28 days after the call was registered. A call is said to have resulted in a request if a request is registered within 28 days after the call is registered. A limited time is necessary for these conversions to make it plausible that the call has directly caused the visit or the request.

To find a reasonable amount of time to wait before it is plausible that a call has nothing to do with the visit or request, the distribution of these waiting times has been taken from the data. These distributions can be seen in Figure 10. On the horizontal axis of these figures the number of days since the call after which a request or a visit was logged given that a request of or a visit was logged. From these two figures it can be seen that after 28 days more than 70% of all requests have been logged. If a much longer period is taken than it is plausible that more contacts than just the call intended for acquisition have taken place. Because of this causality between the call and the visit or the request. In combination with the insights gained from the interviews it can be concluded that 28 days is a sufficiently long time to ensure a dependable conversion but at the same time a valid conversion. Or in other words, that the reliable causality between the call and the visit of the request is plausible.

3.9.2 sources of data
On the basis of the interviews different sources of relevant data sources have been selected. In this part of the report these data sources are listed and described. It is also stated where these sources will be accessed. A further selection based on the feasibility and accessibility will also be described here. The technical discussion about how to use the sets or how to implement them into the model will not be held here but later on in the report.

The list of sources and a description, based on Table 4 is the following:

- BVNL: a list of all the companies active in the Netherlands.
- Jobfeed: A list with all the job listing posted online powered by the company Textkernel.
- Nielsen: a database with job openings that also contains the openings on paper with about 40% of all the openings in the Netherlands. This contains the job listings per industry.
- ROA: A research project of the Maastricht University in collaboration with Randstad that gives insight into the current and future situation on the employment market with the variables. This data set also contains the temporary staffing ratio per industry.
- Purchasing manager’s index: An index that measures the sentiment of purchasing managers about the economy.
- CBS business cycle data (business cycle survey): A source that has the economic prospects per industry.

Note that not all these sources are accessible or can be used for the purpose of this thesis. The BVNL data, which is the total collection of economically active companies in the Netherlands. For the building of the model only internal companies are used since the result of a call is needed in order to do supervised learning. Therefore, the BVNL is only used when validating the model. Jobfeed and Nielsen could be useful sources, like indicated in the interviews. But in order to be useful it has to be connected to the type of function. Given the state of the internal taxonomy of functions this yields a challenge that goes beyond the scope of this thesis. The purchasing manager’s index cannot be accessed so that is the
reason why it cannot be used in the context in this thesis. So, the ROA and CBS business cycle data are used as external data sources for building the model. The BVNL data will be used for validating the model.
4. Data
For the algorithms described in the literature algorithm to function properly the data cannot be used in the initial form. The act of altering and manipulating the data in order to get them ready for the algorithms is called pre-processing. First, an exploration of the data will be given. On the basis of this exploration, it is concluded which different steps are necessary to make reliable predictions. In the context of this thesis, this consists of two steps: techniques on how to use imbalanced data and techniques for using temporal data. These two steps are described in this chapter.

4.1 Exploration of the data
In order to be able to build an analytical model the data first has to be adapted to be interpretable by the model. The high level steps that are taken are described in Figure 11. Below, each of these nodes will be described in detail.

![Diagram](image)

**Figure 11; The pre-processing steps from raw data to pre-processed set.**

4.1.1 Internal Mondriaan data
In Figure 11 the internal data is depicted by the Mondriaan data node. Mondriaan is the name of the internal database. The internal data consists of three elements:
• Contact data: data about the contact that were made with (potential) clients
  o Unique contact ID and ID of the company that has been contacted
  o The type of contact and the intent of the contact. In the case of this thesis this is limited to the type Telephone and the calls with the intention of generating sales.
  o Date at which the contact took place.
• Request data: data about the requests made by companies for temp worker
  o ID for the company that did the request.
  o Information about when the request took place and when the temp worker is needed. This is in the format of a date.
  o Information about the type of temp worker that is needed. This refers to the function that is needed.
• Company data
  o Unique company ID
  o Different level of industry categories of the company. This is provided in four levels. The first level being the less detailed and the fourth level being very described. This description is determined by the Dutch Central Agency for Statistics (CBS, 2014).
  o Information about the size of the company. This is measured in different categories on the basis of the number of people working at the company.
  o Information about the location of the company in the form of a Dutch postal code.

The use of this data is twofold, it is used as features in itself and it is used to connect to external data.

4.1.2 External data
The external data is mainly data with elements of time, place, and industry. As a result of the interview phase several sources for useful external data have been listed. Due to feasibility issues not all of that sources have been used, as has been described in section 3.9.2. For instance, the internal taxonomy of different functions does not map onto externally used taxonomies. This renders all the sources containing function specific data infeasible to use. Of the list from the interview phase two sources have been implemented.

• ROA
  o Portion of temp staffing in industry. This is the average percentage of labor in a specific industry that is fulfilled by temp workers.
  o Sensitivity to business cycle in industry. This is an index number that indicates to what extend an industry is sensitive to changes in a business cycle.
  o Labor market region. The Netherlands divided into different regions based on the law for social security (Regioatlas, 2014).
  o Congregation (“gemeente”). Each of the labor market regions contains a set of congregations.
  o Province
• Business cycle information (CBS). This data is taken from the statistical agency of the Dutch government. For the indexes used here the data is available from 2008 up to the first quarter of 2014.
- Business cycle: prices index. This is an index that indicates the height of prices.
- Business cycle: prices increase. This indicates the growth of the index.
- Business cycle: number of bankrupts. An absolute number about the amount of bankrupts.
- Business cycle: number of bankrupts compared to a year earlier. The growth of the amount of bankrupts as compared to a year earlier.
- Business cycle: revenue index. An index that indicates the business cycle in terms of revenue per sector.
- Business cycle: revenue increase. The increase of the revenue index per sector.

Data mining has mainly focused on data classification, data clustering, and relationship finding. An issue that emerges during this process is the treating of temporal data (Antunes & Oliveira, 2001). The general goal of using temporal data is to enrich the data and to predict future behaviors. In the case of this thesis a classification model is used for prediction. The problem however that needs to be solved is the representation of temporal data and the pre-processing steps that have to be taken to incorporate this temporal data. This problem applied in the area of classification has been researched on a small scale (Antunes & Oliveira, 2001).

The problem of the representation of temporal data can be addressed in several different ways (Antunes & Oliveira, 2001). There are methods that only use minimal transformation to obtain manageable sub-sequences. Another option is to map the data to a more useable space. This means that the data can be mapped into different categories. The data for this thesis is partially supplied in that format and can be used directly. For instance, the ‘portion of temp staffing in industry’ variable, see Table 5, is not a percentage but ‘high’ or ‘low’. Other data that is used for this thesis consists of index numbers. These are not adapted and are used directly. This gives information about the stance of the total size of an effect. For instance, the ‘prices index’ variable as described in Table 5.

The data concerning certain periods of time, for instance the number of bankrupts in a certain quarter in a certain industry, will be linked to calls from that same quarter and from the same industry. It is also possible to use data from previous quarters and from different industries but that quickly becomes infeasible inside the scope of this thesis.

4.1.3 Creating the matched set

In order to connect external data effectively contact data, request data, and company data have to be merged. First, the calls intended for acquisition have to be selected from the contact data. Also, the planned visits have to be selected from the contact data. This is done using the company ID that is available in all three of the types of internal data. This is done in the ‘matching request to contact data’ and the ‘match visit to contact moment’ in Figure 11. The next step is to calculate the conversion. This is done by comparing the date of the call with the date of the request and the data of the planned visit. If either of those is within 28 days of the call, as determined in the definition, then a one is placed in a new column, otherwise a 0.

The final part is about merging the external data to the internal data. This is done using postal code, dates, and industry. As a final step, the rows that contain too much empty values are filtered out.

This yields a final table that is ready for analysis. A small view of the table is provided in Table 5.
In total there 102,130 cases in the above described table. The first column is the unique identifier, the call ID. Call conversion is the last listed variable because it is the target variable. The goal of the model that will be build is to predict for a specific call the value of the call conversion.

The next step is to look at these variables in more detail. The most important variable to look at is the conversion. The average of the conversion is 7,1%, which means that in 7,1% of the 102,130 cases the value of the column conversion is 1. This is a very skewed distribution and that can cause problems for building an effective model (He & Garcia, 2009). This problem is so severe that the next separate section will be devoted to it.

<table>
<thead>
<tr>
<th>Call ID</th>
<th>Company name</th>
<th>Date of call</th>
<th>Visit (Y/N)</th>
<th>Date of visit</th>
<th>Request (Y/N)</th>
<th>Date of request</th>
<th>SBI-segment 1</th>
<th>SBI-segment 2</th>
<th>SBI-segment 3</th>
<th>....</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td>Brouwers &amp; Zn. metselwerken</td>
<td>01-04-2011</td>
<td>Y</td>
<td>15-04-2011</td>
<td>N</td>
<td>08-04-2011</td>
<td>A</td>
<td>14</td>
<td>145</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>SBI – segment 4</td>
<td>Company size</td>
<td>Postal code</td>
<td>Function description</td>
<td>Municipality</td>
<td>Province</td>
<td>Labor market region</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1456</td>
<td>20-50</td>
<td>1234AB</td>
<td>Production</td>
<td>Haarlem</td>
<td>Noord-Holland</td>
<td>Noord-Holland West</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>Portion of temp staffing in industry</td>
<td>Sensitivity to business cycle in industry</td>
<td>Month of the call</td>
<td>Quarter of the call</td>
<td>Year of the call</td>
<td>Business cycle: revenue index</td>
<td>Business cycle: revenue increase</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>1.23</td>
<td>04</td>
<td>2</td>
<td>2011</td>
<td>105</td>
<td>-12%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>Business cycle: prices index</td>
<td>Business cycle: prices increase</td>
<td>Business cycle: number of bankrupts</td>
<td>Business cycle: no. of bankrupts compared to a year earlier</td>
<td>Conversion(1/0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>19%</td>
<td>13156</td>
<td>12</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: An example of the used data.
4.2 Imbalanced data

One of the characteristics that is found when taking the data about the call conversion into consideration is that the two classes of call conversion are severely imbalanced. This problem has been attracting attention in both practice and academia in the past couple of years since it is often encountered in growing areas like direct marketing and fraud detection (He & Garcia, 2009; Shin & Cho, 2003). This kind of imbalance is called between-class imbalance (He & Garcia, 2009). The minority class of call conversion is of an intrinsic nature since it will never be the case that all work in the Netherlands is done through a staffing agency, i.e. the imbalance is a direct result of the underlying nature of the data space (He & Garcia, 2009).

The first problem that the imbalance poses is that performance of commonly used algorithms like decision trees and support vector machines deteriorates (Tang, Zhang, Chawla, & Krasser, 2009). This is a the case because the algorithm needs to outperform the trivial classifier that just labels everything with the majority class (Provost, 2000). Or more precisely put, dataset complexity is the primary determining factor of classification deterioration, which, in turn, is amplified by the addition of a relative imbalance (He & Garcia, 2009). Data complexity is an umbrella term that refers to issues like small disjuncts, lack of representative data, overlapping and others. The complexity that exists mainly comes from the human factor of not implementing data correctly or completely. The second problem that class imbalance causes is that the commonly used measurements of performance for predictive models do not properly assess the effectiveness of predictive algorithms (He & Garcia, 2009; Tang, Zhang, Chawla, & Krasser, 2009).

These two problems will be addressed separately in the remainder of this chapter. Note that it is not the goal to provide a comprehensive overview of all the solutions available in literature but to depict those specific solutions relevant for this thesis.

4.2.1. Two techniques for assessing the imbalanced data problem

For the problem of deteriorating performance of the algorithms generally two solution areas are described in literature: Sampling techniques and cost sensitive methods (He & Garcia, 2009; Tang, Zhang, Chawla, & Krasser, 2009). First the logic behind these techniques is explained. Secondly, out of the many variations of sampling techniques that exist the relevant ones are described in more detail.

4.2.2. Sampling techniques

Sampling techniques rely on altering the imbalanced input data and generating a balanced set. Studies have shown that these different sampling techniques aid in improved classifier accuracy, justifying the use of these techniques (He & Garcia, 2009). There are two main classes of sampling techniques: oversampling and undersampling (Tang, Zhang, Chawla, & Krasser, 2009).

The most straightforward sampling method is random oversampling or random undersampling. These methods randomly sample rows from the minority class and add these to the data set or randomly remove samples from the majority set, respectively (He & Garcia, 2009). Each of these methods has its own drawbacks. For undersampling the problem is obvious to see: it can lead to loss of information and thus the classifier might miss important structures in the majority class. For oversampling, the problem is more delicate. Multiple instances of certain examples become ‘tied’, leading to overfitting (Mease,
Wyner, & Buja, 2007). To overcome these problems more advanced sampling algorithms are devised. One notable algorithm is the ‘synthetic minority oversampling technique (SMOTE), which has shown success in various applications (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). SMOTE is an oversampling approach in which the minority class is oversampled by creating synthetic examples rather than by oversampling with replacement. This is done by taking the k nearest neighbors of each sample of the minority class and introducing synthetic examples the lines joining one or more of the k nearby samples. Dependent of the amount of oversampling desired, the amount of neighbors is selected randomly. For generating the synthetic examples: take the difference between the sample and the selected nearest neighbor and multiply this by a random number between 0 and 1. Then add the result to the original sample. SMOTE has some drawbacks however, including over generalization and variance. These problems stem from the fact that SMOTE generates samples without looking at neighboring samples, thus increasing the occurrence of overlapping between classes. There are more advanced models available in literature that provide limitedly enhanced performance (Tang, Zhang, Chawla, & Krasser, 2009; He & Garcia, 2009). However, since SMOTE provides good performance (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) it is used in this thesis.

4.2.3. Cost sensitive methods
As opposed to changing the distribution of the input dataset, as in in sampling, cost-sensitive methods introduce a penalty for misclassification. This is done by introducing costs associated with misclassification. Various empirical studies have suggested that in some imbalanced learning domains, cost sensitive learning is superior to sampling methods (He & Garcia, 2009).

The basis of the cost-sensitive learning technique is the cost matrix. In this matrix the numerical representation of the penalty per classification is stored. Typically, there is no cost for correct classifications. The goal of cost sensitive methods is to find a model that minimizes the total cost. These cost sensitive methods can be applied to decision trees and support vector machines. In the case of decision trees it can take three forms: cost-sensitive adjustments can be used for the decision threshold, cost-sensitive pruning can be applied to the pruning method, and cost-sensitive considerations can be used for the split criteria. The details of these methods are described extensively in literature (He & Garcia, 2009) and will be used in the context of this thesis. The specific implementation that has been used has been taken from literature (Domingos, 1999).

4.3 Metrics for assessing performance
The usual accuracy metric, based on misclassification, fails in the context of imbalanced data since a trivial classifier that just predicts the majority class performs well under this measure. For adequate measurement of the performance of imbalanced algorithms different metrics are necessary. This has been addressed extensively in the literature (He & Garcia, 2009). Here the relevant metrics for this thesis are described. In Table 6, these metrics are depicted.
The F-measure described in Table 6 combines both recall (completeness) and precision (correctness). The beta can take any non-negative value and is a way to emphasize one of the two. Normally a value of $\beta = 1$ is chosen. A limitation of the F-measure is that it remains sensitive to data distributions (He & Garcia, 2009).

The last one can also be applied to families of classifiers (Menon, Agarwal, & Chawla, 2013). The ROC-graph provides a graphical representation of the trade-offs between the benefits (the true positives), and the costs (the false positives). An example of such a curve can be seen in Figure 13. The line A represents a perfect predictor, the line C the totally random predictor, and the line B a predictor with performance somewhere in between. The value of the area under the curve is also often used as a measure of accuracy and is called the ‘area under curve’. This value varies from 0.5 (totally random) to 1 (perfect predictor) (Baecke & Van den Poel, 2011). The ROC method has also received criticism (Lobo, Alberto, & Real, 2008). The most relevant part of criticism it that the ROC summarizes the performance over different regions of the data.
Apart from the theoretical measures, also a cost measure is used. The measurement that is used is the potential savings stemming from the fact that less calls are needed to achieve the same amount conversions as compared to the original situation. To do this a hypothetical test of 10,000 calls is taken into consideration. This is shown graphically in Figure 12. Under the old conversion of 7% this would yield 700 conversions. Then the amount of calls needed to get this number of conversions under a model is calculated. The model is used on big set of potential calls to generate this subset. To make sure the model is applicable the size of the needed set cannot exceed 60,000 companies. This number has been taken from the business as this the amount of companies that is used to construct the set of companies used for the test in the validation phase. It ensures that the models can be used in further testing. The costs per call are also taken from the business and are €7,- per call. In the case of the old conversion of 7% this would yield a total cost of €70,000. When using a model the number of calls is less than 10,000. The savings are then calculated by taking the difference between 10,000 and the number of calls needed using the model. This number is then multiplied by €7,-.
5. Building the model
As a starting point for the model building the dataset that followed from the pre-processing is used. The different types of models that are used are described in the literature chapter. First, the overview of the different models that will be built can be seen in Table 7. Secondly, the flows that are used to implement the models into the tools that have been used are shown.

<table>
<thead>
<tr>
<th></th>
<th>Decision Tree</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>J.48 implementation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMOTE</td>
<td>Yes/No</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Parameters</td>
<td>Pruning, reduced error pruning, subtree raising, binary splits.</td>
<td>Number of trees to build</td>
</tr>
<tr>
<td>Cost sensitive</td>
<td>Yes/No</td>
<td>Yes/No</td>
</tr>
</tbody>
</table>

Table 7; The different parameters of the model building.

For each of the algorithms several different scenarios will be tested. Each scenario consists of a different combinations of parameter settings, whether SMOTE is used and a setting of cost sensitivity. There are different parameters that can be either turned on or off. For the j.48 decision tree pruning, reduced error pruning, and subtree raising refer to how the trees will be pruned and can all be either turned on or off. Pruning is the process to make large trees smaller in order to reduce the risk of over fitting while not hurting performance too much. When binary splits is turned on every nominal value is split in a binary fashion, i.e. ‘category x’ is the left branch and ‘not category x’ in the right branch. For the random forest, the number of trees to be built can be given a value. Obviously, increasing this value comes at a computational cost. Both of the algorithms are very tolerant as to what data can be entered so no extra pre-processing is needed to make the models work. In the case of SMOTE, a model will be built that uses the SMOTE up sampling technique and one that does not use this technique. For the different parameter settings the results are recorded and the best are selected. For the cost weighing, this is used in every case when it is implemented in the tool.

The models have been implemented in the data mining tool Knime build 2.10. Knime is an open source modular environment, which enables easy visual assembly and interactive execution of a data flow (Berthold, et al., 2009). On top of Knime, the Weka implementation is used since that provides more flexibility regarding the algorithms. (Hall, et al., 2009)

The main flow that is used is shown in Figure 9. In the File Reader node the pre-processed set is loaded into the program. The partitioning divides the data up in two set split 70/30. For implementing the cost sensitivity of the classifiers, the MetaCost node of Weka is used. This is a general method for making classifiers cost sensitive (Domingos, 1999). Inside that node the different models can be selected to use the cost sensitive model on. In the case that no cost sensitivity is desired, the costs inside the MetaCost node are all set equal. The SMOTE node is used for generating a synthetic over sampled set (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Naturally, the SMOTE algorithm is only used on the training data. In
case SMOTE is not needed, the node is circumvented in its entirety. The Weka Predictor node applies the model to the training data. Finally, the Scorer and the ROC curve node provide the necessary results. The scorer node provides the confusion matrix. The ROC curve provides the ROC curve and the area under curve (AUC).

Figure 14; The general flow built in Knime.
6. Results of the models

From the different model that have been built the best instances of every model are described here. The rest can be found in Appendix I: The results from the other models.

Two measurements are used for the selection of the best model: the f-measure and the cost savings measurement. Both those measurements are described in section 4.3. The description of the results will be done in two steps. First, the best models, in terms of the theoretical F-measure and the potential savings are given. This is done for each of the prediction techniques using both the internal and the external data. Secondly, the effect of using external data is described. Apart from the f-measure and the cost savings measurement also three other measurements are given.

The results of the best instance of a model class are shown in Table 8. These results are obtained using both internal and external data (see chapter 4). For the random forest optimized for potential savings both SMOTE as cost sensitivity is used. A false negative has been made twice as ‘expensive’ as a false positive, i.e. a ratio of 2:1. This is much smaller penalty than what was indicated from practice. The number of trees that has been built is 10. The random forest optimized for f-measure does not use SMOTE and has a bigger cost ratio of 100:7. The number of trees is 50. Interesting here is that SMOTE is not used in both cases. This might be explained by the fact that the problem of imbalance is addressed too strong and that performance decreases when a high ratio of cost classification in combination with SMOTE is used. It is interesting to note that the high ratio cost classification performs better when optimizing for f-measure. For the decision tree optimized for potential savings SMOTE was used but no cost savings. Also, reduced error pruning, subtree raising, and binary splits are used. For the decision tree optimized for f-measure both SMOTE and cost sensitivity are used. A cost ratio of 2:1 is used. Again, reduced error pruning, subtree raising, and binary splits are used. Apart from the model, also the current selection method is shown. Obviously, no recall, f-measure and AUC can be calculated. The savings are €0 as this is the situation with which the other models are compared.

The overall impression is that the best random forest has a little better precision at the cost of a little less recall than the best decision tree. This is the case that the two models with the highest F-measure are compared but also when the two models with the highest potential savings are compared. The listed are all feasible so the highest precision yields the greatest savings. So the highest savings come from the random forest optimized for the highest savings yielding €48.355 of savings.

There are two other measures available in the table. The companies needed for set of 700 means is the size of the set that is needed to replicate the conversion of a set of 10,000 calls using no model. Interesting about these results is that the more advanced random forests performs better than the decision tree. This is probably due to the fact that the data contains missing values, which the random forest handles better.

The area under curve (AUC), the measure for how well the classifier performs overall, is clearly higher than 0.5 for both of the classifiers. This means that both the classifiers perform better than random. The random forest has a higher AUC than the decision tree, as can be seen in Table 8. However, because of the imbalanced classes, a higher value of AUC does not directly mean a better classifier.
<table>
<thead>
<tr>
<th>Model</th>
<th>F-measure</th>
<th>Precision</th>
<th>Recall</th>
<th>Companies needed for 700 conversions</th>
<th>F-measure</th>
<th>AUC</th>
<th>Potential savings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random forest</td>
<td>0.17</td>
<td>0.39</td>
<td></td>
<td>4090</td>
<td>0.48</td>
<td>0.69</td>
<td>€41,368</td>
</tr>
<tr>
<td></td>
<td>0.23</td>
<td>0.19</td>
<td></td>
<td>3092</td>
<td>0.40</td>
<td>0.67</td>
<td>€48,355</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.17</td>
<td>0.39</td>
<td></td>
<td>4309</td>
<td>0.48</td>
<td>0.66</td>
<td>€41,764</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.24</td>
<td></td>
<td>5629</td>
<td>0.44</td>
<td>0.63</td>
<td>€45,729</td>
</tr>
<tr>
<td>Current selection method</td>
<td>-</td>
<td>0.07</td>
<td>-</td>
<td>10000</td>
<td>-</td>
<td>-</td>
<td>€0</td>
</tr>
</tbody>
</table>

Table 8; The results of the best models using all data

Figure 15; The decision tree versus the random forest, max costs saved measurement with all the data used.
In Figure 15 and in Figure 16 the resulting ROC-curves are given. The grey lines in these figures indicate a fully random model. Hence, the gray line has an AUC of 0.5. The red lines indicate the performance of the model. The ROC curve is a plot of the true positive rate versus the false positive rate for different thresholds of the model. The further the red line is curved to the top left of the figure, the less false positives are needed to get more true positives. What can be seen is that there is a big difference between the decision tree and the random forest. For both the decision trees it can be seen that for increasing thresholds the tradeoff between true positive rate and false positive rate deteriorates. This is much less so for the random forest.

For the second step, the effect of excluding external data will be examined. In Table 9 the performance is listed. For these results, the same models as in Table 8 are used. The first thing that comes to mind is that the precision only increases by 0.01 by adding the external data. This goes for both the decision tree as the random forest. However, the recall decreases significantly when adding the external data in case of the random forest. This is not the case for the decision tree. This can be explained by the fact that the external data contains more missing values than the internal data. Random forests tend to outperform decision trees in that case (D’Haen, 2013). The regular decision tree using only the internal data performs worse than the decision tree when using all the data. But the difference in performance is smaller than for the random forest. This supports the claim that the external data indeed adds value to the performance. It also supports the claim that the external data contains more missing values because the random forest can handle that better. Hence the bigger performance leap for the random forest.
<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>AUC</th>
<th>Companies needed for 700 conversions</th>
<th>Potential savings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random forest internal data</td>
<td>0.15</td>
<td>0.33</td>
<td>0.40</td>
<td>0.66</td>
<td>3503</td>
<td>€45.472</td>
</tr>
<tr>
<td>Random forest all data</td>
<td>0.17</td>
<td>0.39</td>
<td>0.48</td>
<td>0.69</td>
<td>3092</td>
<td>€48.355</td>
</tr>
<tr>
<td>Decision tree internal data</td>
<td>0.16</td>
<td>0.36</td>
<td>0.44</td>
<td>0.63</td>
<td>3708</td>
<td>€44.127</td>
</tr>
<tr>
<td>Decision tree all data</td>
<td>0.17</td>
<td>0.39</td>
<td>0.48</td>
<td>0.66</td>
<td>5629</td>
<td>€45.729</td>
</tr>
</tbody>
</table>

Table 9: The results for the model using the internal data only compared to using all data

Figure 17: The ROC curves for max cost savings using only internal data

In Figure 17 the ROC curves using only the internal data can be seen. A similar behavior as for the ROC curves with all the data can be seen. This means that the external data does not change the structure of the data in the sense that the random forest still performs better than the decision tree.
7. Validation
Apart from the learning on historical data the model be tested using recent data directly from the field. Firstly, an description of the campaign and how the data was gathered is given. Secondly, the details of the test are given. This includes the different assumptions that are necessary to be able to use the data. Finally, the results of the validation will be stated.

The call campaign has been executed by the in-house call center calling about 4500 companies. The set is divided into about 3379 random companies and 1214 companies selected by the company. These companies all are SME companies of a size between 5 and 50 FTE. Each of these companies has not been a customer in the last two years, the so called cold leads. These leads were all cold called in June and July 2014. These calls enquire whether the company has interest in an appointment with a sales representative of the company. If they do, the call is considered a success i.e. conversion.

Since it has proven infeasible to execute a sales campaign specifically for testing the model the data from this campaign has been used. Because of this, several assumptions have to be made. First, the definition of success changes as compared to the rest of this thesis. In the rest of the thesis the definition that is used is that a call is a success when either an appointment or a request is logged within 28 days after the call. This is infeasible to replicate in the case of this campaign, as the companies that have been called cannot be traced back consistently to the systems where the requests are stored. In the case of this campaign a call is called a success when the company indicates during the call that it is interested in being further contacted. Therefore it is assumed that the definition shows similar behavior to the definition used throughout this thesis. Secondly, the model that has been built uses, among other things, quarterly data about the business cycle. This data has been gathered from the Dutch Central Bureau for Statistics. However, since this campaign has taken place in the second and third quarter of 2014 this data has not been made available yet. To overcome this, the values for these two quarters will be extrapolated based on the historical data. This method is often used in practice (Amstrong, 2001). To adjust for seasonal effects that these numbers have the average of the differences of the previous years will be added to the value of the first quarter of 2014.

Let $BCN_{i,j,k}$ be the estimated Business Cycle Number $i$ of quarter $j$ in year $k$.

Then for the estimate of the second quarter for the business cycle number $i$.

$$BCN_{i,2,2014} = BCN_{i,1,2014} + \frac{\sum_{k=2008}^{2013}(BCN_{i,2,k} - BCN_{i,1,k})}{k}$$

And for the third quarter the estimate is given by:

$$BCN_{i,3,2014} = BCN_{i,1,2014} + \frac{\sum_{k=2008}^{2013}(BCN_{i,3,k} - BCN_{i,1,k})}{k}$$
7.1 The results of the validation

For validating the model a comparison between the performance of the randomly selected set of 3379, the set predicted by the company, and the set of the companies selected by the model is made. The model will be executed on the set of 1214 companies that have been selected by the company. Then, from the companies that were predicted as a conversion, the proportion of correctly predicted companies is calculated. This is equal to the precision, as has been used earlier.

The results of the validation are displayed in Table 10. The random set is the control group to which the other two are compared.

<table>
<thead>
<tr>
<th></th>
<th>Number of cases</th>
<th>Conversion</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random set</td>
<td>3379</td>
<td>7,9%</td>
<td>-</td>
</tr>
<tr>
<td>Set selected by the company</td>
<td>1214</td>
<td>13,3%</td>
<td>68%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>1214</td>
<td>11,3%</td>
<td>43%</td>
</tr>
<tr>
<td>Random Forest</td>
<td></td>
<td>7,9%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 10: The results of the validation

The results for this data are seemingly worse than the results achieved in the results phase. However, this is not necessarily the case as the conversion that is displayed is not the same conversion as has been displayed in the rest of the thesis so these numbers cannot be compared one to one. As can be seen the decision tree performs better than the random forest. This can be an indication that the results cannot be generalized very well since random forest tend to outperform decision trees in general. Especially with respect to generalizability. The model is not trained on this data so over fitting is not the case but it might be the case that the decision tree that has been used works especially well with this data. The performance is also likely to be damaged by the different assumptions that have been made to produce these results: The extrapolation of the data and the different definition of conversion that is used.

Another thing that potentially hurts the performance is that the used from the call action only consists of the SME companies but the model is built for general purpose. If a model is built specifically for this case the performance is likely to be better. Another potential problem with this validation is that the set on which the model has been ran on a set of companies that already was a preselected by Randstad. Because of this it is not an unbiased test. In this validation the companies selected as ‘conversion’ if they are predicted by both Randstad and the model built for this thesis. It might be the case that the model built for this thesis would have correctly predicted companies that are not in the preselected set of Randstad to begin with.
8. Conclusions
The conclusion of this thesis will consist of three parts. In the first part the research questions and the overall goal of this research will again be taken into consideration. This will be described in section 8.1. The second part reflects on the limitations of the research and is depicted in section 8.2. The third and last part provides options for further research in section 8.3.

8.1 Revisiting the research questions and the overall goal
This section is dedicated to revisiting the research questions and answering them, thus drawing conclusions and going towards the outcome of this thesis.

8.1.1 Suitable data sources for prediction and a definition of call conversion
Research question one, which consists of two sub-questions, was defined in section 1.4 as follows:

Research question 1
What data sources are suitable in the context of predicting the call conversion?
1a: What external data sources can be used to predict the value of the call conversion?
1b: What is a usable definition for call conversion and how is this ‘datafied’ internally?

These two questions have been answered using an interview approach as has been elaborated upon in chapter 3. For these interviews people from the business departments and the IT department ensuring a balanced view.
In a total of four interviews a list of as many possible different potentially useful sources of external data were determined. Subsequently, taking feasibility into account, this list was shortened to the data sources that have actually been used in the building of the model.
This yielded the following two sources:

- ROA: A research project of the Maastricht University in collaboration with Randstad that gives insight into the current and future situation on the employment market with the variables. This data set also contains the temporary staffing ratio per industry.
- Purchasing manager’s index: An index that measures the sentiment of purchasing managers about the economy.

From each of these sources different variables were taken to be used in the building of the model. Thus research question 1a is answered.
The second part of research question one is about the definition of call conversion. This has also been answered using the results of the four interviews.
Taking the internal data into consideration the used definition for call conversion becomes the following.

‘Call conversion is defined in practice as the percentage of outbound calls to targeted companies (either suspective, prospective or inactive companies), intended for acquisition, that takes a targeted company further into the sales funnel by either a visit or a staffing request within 28
This definition can be measured using the internally available data and thus research question 1b has been answered.

8.1.2 The best predictive model to predict call conversion

In this section the second research question is answered. This was defined in section 1.4 as follows:

**Research question 2:**
What is the best predictive model that can be built to predict call conversion using the results of RQ1?

This question has been approached in two steps. First, two potentially useful models have been taken from literature as has been described in chapter 2. Two types of machine learning models have been used: the decision tree and the random forest.

Using the internal historical data the definition of conversion has been implemented. Subsequently, the external data has been connected to this internal data. This process has been described in chapter 4. For building the best model different parameters have been optimized as has been described in chapter 5. One of the main problems to solve for this question is to define what the best model is. This is challenging the conversion data was heavily imbalanced. This made the regular accuracy statistic unfit for determining performance, a problem that has been extensively discussed in literature. To assess model performance the theoretical f-measure and a specially constructed cost savings measure have been used. This has been elaborated on in section 4.2.4. To answer the second research question the best models per technique and per performance measure are listed below in Table 11.

<table>
<thead>
<tr>
<th></th>
<th>What is the highest</th>
<th>F-measure</th>
<th>Potential savings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random forest</strong></td>
<td>F-measure</td>
<td>0,48</td>
<td>€41.368</td>
</tr>
<tr>
<td></td>
<td>Potential savings</td>
<td>0,40</td>
<td>€48.355</td>
</tr>
<tr>
<td><strong>Decision tree</strong></td>
<td>F-measure</td>
<td>0,48</td>
<td>€41.764</td>
</tr>
<tr>
<td></td>
<td>Potential savings</td>
<td>0,44</td>
<td>€45.729</td>
</tr>
</tbody>
</table>

*Table 11; The best models per technique and per performance measure*

Note that these results do not guarantee that this is the best possible model that can be built for this data. Because the limited timeframe of this thesis project, not all possible models but only two types have been tried.

8.1.3 The effect of using the predictive model on the business in practice

In this section the third and last research question is answered. This research question was stated in section 1.4 as follows:
In order to address this question the model resulting from research question 2 have been taken and applied to a different validation data set. This data set has been taken from a call action done by the Randstad in-house call center. The list of companies consisted of SME companies that were not customers of Randstad. Half of the set consists of randomly selected companies and the rest was selected by Randstad. One problem was that the definition of call conversion used throughout this thesis cannot be used for this dataset. However, similar conversion data is available and will be used. Also, since the call action has been executed in the second and third quarter of 2014 not all the external data was made available yet. To overcome this, this data has been constructed by extrapolation. This has been described in chapter 7. The results of the validation are given below in Table 12.

<table>
<thead>
<tr>
<th>Number of cases</th>
<th>Conversion</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random set</td>
<td>3379</td>
<td>7,9%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>1214</td>
<td>11,3%</td>
</tr>
<tr>
<td>Random Forest</td>
<td></td>
<td>7,9%</td>
</tr>
</tbody>
</table>

Table 12; The results of the validation.

As can be seen the decision tree performs better than the random forest and the random set. This indicates that the model is to some extend robust to estimated data and approximate definitions. This answer is similar to the answer to research question 2 in the sense that it shows that a model works better than random. It differs from that answer however because it shows strong indications of robustness and good performance in practical application. One of the other possible reasons why the model did not perform better might be that the model built in this thesis is a general model with the respect to company size. However, the companies called in this call action were all SME companies. This can hurt the performance.

8.1.4 Reflection of the overall goal and the overall contribution of this thesis

In this section a reflection is given on the overall goal which was presented in section 1.4 as follows:

**Overall goal**

_Edward McMillan’s goal was to predict call conversion using external data sources on top of the traditional internal data sources. Secondly, implement this prediction model to enhance business performance._

In the previous sections of this chapter the different intermediary goals have been described. Overall it can be said that the goal of this thesis was achieved. A model was built that used external data in addition to the internal data and this model was able to predict call conversion better than without using this model. Also, this was put to the test in a validation using data from a call action. Despite several assumptions that had to be made to be able to use the data to test the model, there still
was an improvement in the performance in that specific dataset. It needs to be said that this is not actual implementation in the strict sense of the word but there is strong indication that it would pose value in that case.

One of the big parts that was not firmly achieved in this thesis was that the external data adds predictive power to the model. This will be detailed in the next section.

8.2 Limitations of the research
In this section the limitations of the research are discussed. This is done by addressing the limitations for each of the parts of the research design as discussed in section 1.4. The limitations of the literature study as discussed in chapter 2 will be discussed in section 8.2.1. The limitations of the in depth interviews from chapter 3, are described in section 8.2.2. The limitations of the building of the model from chapter 4 through 6, are described in section 8.2.3. Finally, the limitations of the validation phase from chapter 7, are described in section 8.2.4.

Overall the biggest limitation in this research has been the intrinsically limited data quality. This stems from the fact that the majority of the internal data used has been manually entered by hundreds of different interagents. There is no incentive in place stimulate correctly data entry thus the data quality remains questionable. Furthermore, the organization is not very mature with respect to unlocking the value from the vast amount of data available. This research is therefore a first step in reaching that final goal.

8.2.1 Limitations of the literature study
In chapter 2 the description of the literature that has been found with regard to the different techniques. One of the main limitations of the literature study is that no formal framework was followed. Therefore, neither completeness as adequate relevance are ensured. However, the specific part of the machine learning field in literature has been searched is relatively small. So it is reasonable to assume that the most influential articles have been found.

8.2.2 Limitations of the in depth interviews
For the interviews an in depth interview approach has been used. There are typically several limitations to this approach (Boyce & Palena, 2006): prone to bias, can be time-intensive, interviewer must be appropriately trained, and it is not generalizable. Especially the first and the third limitations might be relevant. Obviously, it is very hard for the interviewer to determine a bias in the interviews. Therefore, it is hard to assess the magnitude of this limitation. It is however worth noting that this might pose a limitation of the outcome of the interview phase. Another limitation is that the interviewer has not received formal training in conducting in depth interviews. Thus it is questionable that the best possible result has been gathered from the interviews.

Another potential limitation is that the amount of interviews is too small. That would mean that not a complete view from the organization has been gathered. No indications that this might be the case have been found so in that case it would be an unknown unknown.

8.2.3 Limitations of the model building
Several limitations to the model building can be identified. First, the problem of selecting which model to use from the vast amount of models available from literature has been solved by looking at similar
studies. However, this does not guarantee that the best model for this specific problem has been used. If this is the case this may place a limitation of the quality of the model. A choice that has been made is to keep the model general. In the sense that all the calls that comply with the definition of call conversion have been used to build the model. Another problem is that the used hardware, in combination with the used tools and size of the datasets provides a limit to the number of iterations that can be done. In the case of the random forest, this would mean that better performance might be possible if more computing power would be available. Another limitation of the research is that the cost savings measurement as used might be an underestimation of the real value to be gained from implementing the model. The real value might be in the fact that more customers are attracted, not in the fact that less calls have to be made. This has not been chosen as a measurement because it requires assumptions about the total number of companies that are interested in staffing services.

8.2.4 Limitations of the validation phase
In the validation phase several potential limitations can be identified. The main limitation is that the call action has not specifically been designed for this thesis. Therefore, it is not necessarily the case that the results will generalize in other call actions. Because of the fact that data has been taken from another call action several assumptions had to be made. In case these assumptions do not hold this severely limits the validity of this validation. The second limitation is that the general model that has been built is tested in a very specific niche of the market. This limits the performance for this specific validation.

8.2.5 Generalizability of the model
When doing research in industrial engineering, often the dilemma between rigor and relevance comes into play. If one focuses too much on the rigor of normal science, the risk exists that the research will not be relevant for practitioners that seek applicable knowledge. If the focus is too much on relevance, the risk exists that the standards of falling short of prevailing standards of rigor (Argyris, 1999). This study has been focused on the relevance aspect as the research was conducted from within a company with the goal to enhance the sales process. Because of this focus the rigor aspect has automatically received less focus. However, the balance between the two is still maintained at an adequate level: the claims and methods that the research is built all come from literature. Because of the focus on relevance the question always rises how well the research generalizes to other instances. When looking at the generalizability the very models are of lesser importance as these have been built specifically for this company using very specific data. It has been shown in the validation that the model still performs reasonably well when applied at different sales data. However, the generalizability of the approach that is used, using predictive modeling to look into enhancing business processes, is interesting. In the case of this thesis two different types of generalizability are relevant. How well the approach generalizes to other processes within the company and how well this approach generalizes to the sales process of other industries. When looking throughout the sales funnel in the company, as can be seen in Figure 9, many more conversion points can be defined and measured. On the basis of that it is reasonable to assume that the same approach as in this thesis can be used throughout the company. This does not imply that the very models used can be applied in different industries that want to predict sales since this model has been specifically built for the staffing industry. Many other industries have similar sales
funnels as the staffing industry and also collect huge amounts of data. Therefore, it can be expected however that the approach for building used throughout this thesis is applicable in other industries. This can be interpreted as a form of generalizability.

8.3 Further research
This section concludes the thesis by indicating what further research is necessary to cover the limitations from section 8.2. In general it is interesting to look into applying predictive modeling to other parts of the sales funnel. This especially becomes the case when the organization matures in using data and the means to extract, preprocess and model the data become readily available to use.

The literature study done in chapter 2 has not been done in a formal systematic way. To improve this a systematic approach can be taken from literature. In that case the completeness can be guaranteed and a much stronger selection of which algorithms to use for building the model can be made.

The interviews of chapter 3 can be improved by getting a formal training in doing these in depth interviews. As a result the questions that are asked would improve. This might increase the effectiveness and decrease the bias associated with these type of interviews. In addition to this more experts might be interviewed to be able to draw stronger conclusions. However, this is not seen as a vital part of the research.

Chapter 4, chapter 5, and chapter 6 could be spun off into independent research. First, the two models that have been used in this thesis might be not the best techniques for the specific characteristics of the data that is used. So it useful to do further research and test more different models in a comprehensive way.

Second, the use of external data has not been elaborated upon extensively in this thesis. There are several ways in which the external data can be researched. First, more data sources can be acquired. Second, more combinations within the data can be made. For instance, the business cycle of earlier quarters instead of the current quarters can be used. This has been attempted superficially in the context of this thesis, see appendix 1.1. It did not show immediate promise so further research in that direction is necessary. In order to do this it might be worthwhile to look at building an (semi-)automated way to do a grid search through the different possibilities.

The validation as described in chapter 7 can be improved by doing a call action dedicated to this model. In that case less assumptions have to be made making the conclusions drawn from this validation stronger. Not only cost savings but also extra customers could be discovered in that case.
Bibliography


http://www.regioatlas.nl/indelingen/indelingen_indeling/t/arbeidsmarktregio_s


Appendices

Appendix I: The results from the other models
In the chapter 6 the results of the best models are described. Here the results of the other models are given. The different columns describe what the setting for that specific run is. This is shown in Table 13 for the decision tree and in Table 15 for the random forest. Both these have a number for the run. In Table 14, for the decision tree, and in Table 16, for the random forest, the results are shown. The ROC curves are not shown.

<table>
<thead>
<tr>
<th>SMOTE</th>
<th>Cost sensitive</th>
<th>Cost matrix</th>
<th>Unpruned</th>
<th>Reduced Error</th>
<th>Pruning</th>
<th>Subtree raising</th>
<th>Binary splits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>No</td>
<td>0</td>
<td>1</td>
<td>FALSE</td>
<td>TRUE</td>
<td>FALSE</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>No</td>
<td>0</td>
<td>1</td>
<td>TRUE</td>
<td>FALSE</td>
<td>TRUE</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>No</td>
<td>0</td>
<td>1</td>
<td>FALSE</td>
<td>TRUE</td>
<td>FALSE</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>No</td>
<td>0</td>
<td>1</td>
<td>FALSE</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>5</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
<td>7</td>
<td>FALSE</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>6</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
<td>1</td>
<td>FALSE</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
<td>1</td>
<td>FALSE</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>8</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
<td>1</td>
<td>FALSE</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>9</td>
<td>Yes</td>
<td>No</td>
<td>0</td>
<td>1</td>
<td>FALSE</td>
<td>FALSE</td>
<td>TRUE</td>
</tr>
<tr>
<td>10</td>
<td>Yes</td>
<td>No</td>
<td>0</td>
<td>1</td>
<td>TRUE</td>
<td>FALSE</td>
<td>TRUE</td>
</tr>
<tr>
<td>11</td>
<td>Yes</td>
<td>No</td>
<td>0</td>
<td>1</td>
<td>FALSE</td>
<td>FALSE</td>
<td>TRUE</td>
</tr>
<tr>
<td>12</td>
<td>No</td>
<td>No</td>
<td>0</td>
<td>1</td>
<td>FALSE</td>
<td>TRUE</td>
<td>FALSE</td>
</tr>
<tr>
<td>13</td>
<td>No</td>
<td>No</td>
<td>0</td>
<td>1</td>
<td>FALSE</td>
<td>FALSE</td>
<td>TRUE</td>
</tr>
<tr>
<td>14</td>
<td>No</td>
<td>No</td>
<td>0</td>
<td>1</td>
<td>TRUE</td>
<td>FALSE</td>
<td>TRUE</td>
</tr>
<tr>
<td>15</td>
<td>No</td>
<td>No</td>
<td>0</td>
<td>1</td>
<td>FALSE</td>
<td>FALSE</td>
<td>TRUE</td>
</tr>
<tr>
<td>Model number</td>
<td>Area under curve</td>
<td>Precision</td>
<td>Recall</td>
<td>Size of positive predictions needed to keep conversion 7%</td>
<td>Set needed to make that many positive predictions</td>
<td>Expected savings</td>
<td>F-measure</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------</td>
<td>-----------</td>
<td>--------</td>
<td>-----------------------------------------------------------</td>
<td>--------------------------------------------------</td>
<td>-----------------</td>
<td>----------</td>
</tr>
<tr>
<td>1</td>
<td>0,605</td>
<td>0,19</td>
<td>0,26</td>
<td>3747</td>
<td>37960</td>
<td>€43.774</td>
<td>0,43</td>
</tr>
<tr>
<td>2</td>
<td>0,62</td>
<td>0,19</td>
<td>0,30</td>
<td>3779</td>
<td>31540</td>
<td>€43.547</td>
<td>0,46</td>
</tr>
<tr>
<td>3</td>
<td>0,612</td>
<td>0,19</td>
<td>0,29</td>
<td>3634</td>
<td>33047</td>
<td>€44.564</td>
<td>0,46</td>
</tr>
<tr>
<td>4</td>
<td>0,621</td>
<td>0,20</td>
<td>0,26</td>
<td>3512</td>
<td>37042</td>
<td>€45.415</td>
<td>0,45</td>
</tr>
<tr>
<td>5</td>
<td>0,541</td>
<td>0,07</td>
<td>0,99</td>
<td>9356</td>
<td>9519</td>
<td>€4.506</td>
<td>0,28</td>
</tr>
<tr>
<td>6</td>
<td>0,645</td>
<td>0,17</td>
<td>0,40</td>
<td>4038</td>
<td>23543</td>
<td>€41.735</td>
<td>0,49</td>
</tr>
<tr>
<td>7</td>
<td>0,668</td>
<td>0,14</td>
<td>0,56</td>
<td>5143</td>
<td>16821</td>
<td>€33.997</td>
<td>0,44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0,15</td>
<td>0,48</td>
<td>4717</td>
<td>19731</td>
<td>€ 36.980</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0,575</td>
<td>0,17</td>
<td>0,24</td>
<td>4131</td>
<td>41645</td>
<td>€ 41.085</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0,577</td>
<td>0,18</td>
<td>0,24</td>
<td>3988</td>
<td>41726</td>
<td>€ 42.087</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0,576</td>
<td>0,17</td>
<td>0,22</td>
<td>4028</td>
<td>45152</td>
<td>€ 41.807</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0,533</td>
<td>0,46</td>
<td>0,02</td>
<td>1532</td>
<td>579657</td>
<td>infeasible</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0,552</td>
<td>0,32</td>
<td>0,05</td>
<td>2159</td>
<td>180229</td>
<td>infeasible</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0,571</td>
<td>0,33</td>
<td>0,04</td>
<td>2130</td>
<td>233123</td>
<td>infeasible</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0,592</td>
<td>0,34</td>
<td>0,05</td>
<td>2038</td>
<td>212350</td>
<td>infeasible</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0,08</td>
<td>0,95</td>
<td>9246</td>
<td>10386</td>
<td>€ 5.280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0,657</td>
<td>0,14</td>
<td>0,52</td>
<td>5115</td>
<td>18797</td>
<td>€ 34.193</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0,642</td>
<td>0,09</td>
<td>0,81</td>
<td>7559</td>
<td>12069</td>
<td>€ 17.090</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0,659</td>
<td>0,14</td>
<td>0,52</td>
<td>5134</td>
<td>19014</td>
<td>€ 34.062</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0,658</td>
<td>0,11</td>
<td>0,70</td>
<td>6541</td>
<td>14018</td>
<td>€ 24.212</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>0,66</td>
<td>0,16</td>
<td>0,45</td>
<td>4309</td>
<td>20802</td>
<td>€ 39.835</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0,657</td>
<td>0,12</td>
<td>0,60</td>
<td>5630</td>
<td>16285</td>
<td>€ 30.592</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>0,625</td>
<td>0,10</td>
<td>0,69</td>
<td>6994</td>
<td>14346</td>
<td>€ 21.043</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0,551</td>
<td>0,08</td>
<td>0,95</td>
<td>9299</td>
<td>10331</td>
<td>€ 4.907</td>
<td></td>
</tr>
</tbody>
</table>

Table 14: The results of the decision tree
<table>
<thead>
<tr>
<th>Model number</th>
<th>SMOTE sensitive</th>
<th>Cost sensitive</th>
<th>Area under curve</th>
<th>Precision</th>
<th>Recall</th>
<th>Size of positive predictions needed to keep conversion 7%</th>
<th>Set needed to make that many positive predictions</th>
<th>Expected savings</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>No</td>
<td>0.605</td>
<td>0.19</td>
<td>0.26</td>
<td>3747</td>
<td>€ 43.774</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>No</td>
<td>0.62</td>
<td>0.19</td>
<td>0.30</td>
<td>3779</td>
<td>€ 43.547</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>No</td>
<td>0.612</td>
<td>0.19</td>
<td>0.29</td>
<td>3634</td>
<td>€ 44.564</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>No</td>
<td>0.621</td>
<td>0.20</td>
<td>0.26</td>
<td>3512</td>
<td>€ 45.415</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Yes</td>
<td>Yes</td>
<td>0.541</td>
<td>0.07</td>
<td>0.99</td>
<td>9356</td>
<td>€ 4.506</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Yes</td>
<td>yes</td>
<td>0.645</td>
<td>0.17</td>
<td>0.40</td>
<td>4038</td>
<td>€ 41.735</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Yes</td>
<td>0.668</td>
<td>0.14</td>
<td>0.56</td>
<td>5143</td>
<td>€ 33.997</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Yes</td>
<td>Yes</td>
<td>0.15</td>
<td>0.48</td>
<td>0.48</td>
<td>4717</td>
<td>€ 36.980</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Yes</td>
<td>No</td>
<td>0.575</td>
<td>0.17</td>
<td>0.24</td>
<td>4131</td>
<td>€ 41.085</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Yes</td>
<td>No</td>
<td>0.577</td>
<td>0.18</td>
<td>0.24</td>
<td>3988</td>
<td>€ 42.087</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Yes</td>
<td>No</td>
<td>0.576</td>
<td>0.17</td>
<td>0.22</td>
<td>4028</td>
<td>€ 41.807</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>No</td>
<td>No</td>
<td>0.533</td>
<td>0.46</td>
<td>0.02</td>
<td>1532</td>
<td>infeasible</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>No</td>
<td>No</td>
<td>0.552</td>
<td>0.32</td>
<td>0.05</td>
<td>2159</td>
<td>infeasible</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>No</td>
<td>No</td>
<td>0.571</td>
<td>0.33</td>
<td>0.04</td>
<td>2130</td>
<td>infeasible</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>No</td>
<td>No</td>
<td>0.592</td>
<td>0.34</td>
<td>0.05</td>
<td>2038</td>
<td>infeasible</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>No</td>
<td>Yes</td>
<td>0.08</td>
<td>0.95</td>
<td>0.95</td>
<td>9246</td>
<td>€ 5.280</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>No</td>
<td>Yes</td>
<td>0.657</td>
<td>0.14</td>
<td>0.52</td>
<td>5115</td>
<td>€ 34.193</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>Model number</td>
<td>Area under curve</td>
<td>Precision</td>
<td>Recall</td>
<td>Size of positive predictions needed to keep conversion 7%</td>
<td>Set needed to make that many positive predictions</td>
<td>Expected savings</td>
<td>F-measure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------</td>
<td>-----------</td>
<td>--------</td>
<td>----------------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>-----------------</td>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0,6573</td>
<td>0,31</td>
<td>0,07</td>
<td>2294</td>
<td>Infeasible</td>
<td></td>
<td>0,24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0,37</td>
<td>0,07</td>
<td></td>
<td>1898</td>
<td>Infeasible</td>
<td></td>
<td>0,23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0,675</td>
<td>0,32</td>
<td>0,07</td>
<td>2213</td>
<td>Infeasible</td>
<td></td>
<td>0,22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0,684</td>
<td>0,22</td>
<td>0,15</td>
<td>3129</td>
<td>Infeasible</td>
<td></td>
<td>0,37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0,671</td>
<td>0,22</td>
<td>0,16</td>
<td>3254</td>
<td>Infeasible</td>
<td></td>
<td>0,36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0,694</td>
<td>0,23</td>
<td>0,15</td>
<td>3062</td>
<td>Infeasible</td>
<td></td>
<td>0,37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0,682</td>
<td>0,16</td>
<td>0,38</td>
<td>4435</td>
<td>€</td>
<td>38.957</td>
<td>0,44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0,69</td>
<td>0,16</td>
<td>0,38</td>
<td>4400</td>
<td>€</td>
<td>39.200</td>
<td>0,45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0,68</td>
<td>0,23</td>
<td>0,20</td>
<td>3048</td>
<td>€</td>
<td>48.662</td>
<td>0,43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0,692</td>
<td>0,17</td>
<td>0,40</td>
<td>4179</td>
<td>€</td>
<td>40.747</td>
<td>0,47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 15: The settings of all the models that have been tested for the random forest.
Appendix 1.1 The results other external data use

It has been tested whether the use of data from earlier quarters instead of the current quarter provides better results. This has been done in a straightforward way by adding the business cycle data. Then the best random forest model has been rerun with this data. This yields the results in Table 17.

<table>
<thead>
<tr>
<th>Area under curve</th>
<th>Precision</th>
<th>Recall</th>
<th>Size of positive predictions needed to keep conversion 7%</th>
<th>Set needed to make that many positive predictions</th>
<th>Expected savings</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.662</td>
<td>0.23</td>
<td>0.20</td>
<td>3042</td>
<td>46123</td>
<td>€ 48.703</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Table 17; The results of the earlier external data

When this results are compared to the model without this extra data (model number 9 in Table 16) it can be seen that there is very little extra savings.
Appendix II: the transcripts of the interview
Here the transcriptions of the interviews are displayed. The transcriptions are in Dutch.

**The Marketing database manager.**
S = Sander J= the marketing database manager.
S: Hij neemt op. Als het uitgewerkt is krijg je het nog een keer terug te zien. Dan kun je wijzigingen aanbrengen.
J: Wachten we rustig af. Doel van het gesprek?
S: Ik zit nu in de interview fase van mijn project. Wat ik moet doen is het vinden van een definitie van call conversie en hypothesen over welke data interessant is. Zo van: ‘Ik zou die data in die branche eens bekijken. Laten we beginnen met de definitie. Tot nu toe heb ik van Peter (Zeinstra red.) een hoop data gekregen en het belangrijkste ding wat ik heb zijn calls gelogd onder werving/acquisitie. De vraag is nu: wanneer is zo’n call nou slaagd/conversie gepleegd? Ik heb alle aanvragen in de tijd en gedept. Zo ja, dan is de call geconverteerd. Ook heb ik alle plaatsingen in de tijd. Als er twee weken na de call een plaatsing gelogd is is dat een ‘extra-conversie’. Die geef ik dan twee punten en ik geef hem een punt in het eerste geval.
J: Welke bron? Datawarehouse?
S: Ja, Peter heeft hem rechtstreeks uit het data warehouse getrokken.
J: Wat ik interessant vind, je hebt over call naar aanvraag en call naar plaatsing. Call naar bezoek neem je niet mee?
S: Zou kunnen, nog niet gedaan, die data heb ik ook.
J: Calls naar recalls?
S: Dat is vrij moeilijk. Die worden niet als dusdanig gelogd.
J: Die kun je mappen mby PGB ID.
Ik ben het met je eens, je kunt het zo doen. Maar in de praktijk zien we het heel veel gebeuren dat er stappen worden overgeslagen. Dat er van een call rechtstreeks een aanvraag wordt. Een plaatsing zal niet veel gebeuren. Vaak komt uit een call een bezoek. Tegenwoordig moet je als je een plaatsing hebt een aanvraag aanmaken in het systeem. Vroeger kon je zomaar een plaatsing genereren. Ik denk dat bezoek en recall belangrijk zijn: ik denk dat het bellen en contact houden met een klant op termijn aanvragen oplevert. Bij staffing is het heel moeilijk om de behoefte te creëren dus je zal altijd de gesprekken aan moeten gaan en de gunfactor moeten krijgen. Ik denk dat de directe ratio call/aanvraag en call/plaatsing relatief laag zijn. Ik denk dat call/bezoek beter is. Bij bel evenementen zien we dat er call/bezoek ratio van 3:1 gehaald wordt. We verkopen geen televisies.
S: Je kunt toch wel betere mensen hebben of goedkoper zijn?
S: Dat gebeurt niet aan de hand van een call
S: Waar ik naar aan het kijken ben om de voorspelbaarheid hoog te maken is dat als ik zo’n genuanceerde definitie ga hanteren dat het vrij moeilijk wordt om te voorspellen. Als ik me de vraag stel: wanneer is een call goed geweest? In principe als er op een redelijke termijn een volgend contact is geweest. Dat is in mijn ogen de maat voor conversie. Dat is natuurlijk te vaag, nu moet ik gaan kijken hoe ik dat precies ga invullen.
J: Je zou met wegen kunnen gaan werken ofzo.
S: Wat je wel ziet is bijvoorbeeld dat er plaatsingen voortkomen uit een call bij een bestaande klant.
J: Ja.
S: Daarom heb ik die wel wat meegenomen. Want er wordt veel toch onder acquisitie werven gelogd. Het is vrij lastig om met het top-of-mind verhaal te werken, aangezien je dan verder in de toekomst aan het voorspellen bent wat het moeilijker maakt. Daarom wil ik mijn definitie ‘dichtbij houden’.
J: (Wijst naar de linkerkant van de vlinder) Dit is de reguliere sales funnel zoals we die hebben. Wat je nu heel terecht ziet is dat een aanvraag uit een call 1 punt waard is en een plaatsing uit een call 2 punten. Dit kun je natuurlijk ook verder opbreken met bijvoorbeeld een recall die een
half punt waard is, een aanvraag twee en een plaasting vijf. Uiteindelijk kun je dan deze parameters in je model gaan aanpassen om te kijken wat het beste werkt.

S: De vraag is of de zwakte van het model zit in deze parameters zit of in de kwaliteit van de data.

J: Klopt, maar ik verwacht omdat dit het sales proces: we bellen om bezoeken te genereren en we bezoeken om aanvragen binnen te halen. Tussen de aanvraag en de plaatsing zit voorziening, wat een onzekerere factor is omdat het met de andere kant van de vlinder te maken heeft. Die doet afbraak aan je scoring. Maar je wilt hem wel meenemen.

S: Uhm, die kan ik op twee manieren meenemen: ik ga werken met een scoring in de definitie van conversie of ik ga hem behandelen als externe bron. Ik weet niet wat de beste keuze is.

J: Als je naar buiten kijkt, dan kijk je naar de markt, dan neem je het feit niet mee waar wij goed in zijn. Met de voorzieningspercentages zou je dat kunnen doen.

S: Als ik bezoek, aanvraag en recall pak, denk jij dan dat ik een goede definitie te pakken heb?

J: Vanuit het reguliere sales process, ja. Er zijn echter nog andere dingen die relevant zijn, bijvoorbeeld zij-instroom.

S: wat is dat precies?

J: Zij-instroom is als we gebeld worden: kom eens langs of ‘hier een aanvraag’. Iedereen kent randstad. Het is op het moment onduidelijk hoe groot het effect hiervan precies is. Het reguliere salesprocess is top down de funnel door met de verschillende conversie ratios. Als we maar genoeg gaan bellen komt er vanzelf wat uit. Je bent eigenlijk op zoek naar de optimale conversie. Deze ga je waarschijnlijk voorspellen aan de hand van drie assen: Branches, functies en geografisch. De zij-instroom is interessant omdat het iets zegt binnen welke branch/regio/functie er opportunities liggen waar we zelf nog niet genoeg op acteren.

S: Eigenlijk is de zij-instroom een soort externe databron met gegevens over de markt die je meet adhv de zij-instroom.

J: Je hebt push en pull. Door te bellen pushen we de markt om aanvragen te doen. Door above-the-line te communiceren pullen we de markt. Dit doen we dmv adwords, merknaam. Zo hebben we ook formulieren op de website met ‘plaats uw vacature gratis’. Als dat gaat groeien en er blijkt een bepaalde branch tussen te zitten die we nog goed kunnen voorzien ook.

S: Dat klinkt als een mooie bron.

J: Inderdaad. Is niet van belang voor je definitie maar handig als externe bron. Als je kijkt naar conversie call-plaastingen geeft een ander beeld dan bezoeken-plaastingen. Als er uit je recall een aanneming voorkomt of het eerste telefoontje zinvol is geweest?

S: Inderdaad, andersom kan natuurlijk ook. Als er uit een telefoontje een recall komt en uit die recall komt vervolgens niks. Is het eerste telefoontje dan een succes geweest?

J: Ja, moeilijk. Maar is wel te achterhalen. Je kunt gaan kijken naar de calls. Ik ga mezelf niet houden aan de reguliere vlinder houden maar ik ga er een laag tussen zetten. Je kunt gaan kijken of er een verband is tussen bedrijven die twee keer gebeld zijn en bedrijven die een aanvraag doen. Als het een niks oplevert weet je dat je er niet naar hoeft te kijken. Als het wel iets oplevert is het heel waardevolle informatie.

S: Dat is een goeie. Wat zeg jij op basis van ervaring dat het meeste voorkomt: dat ze mooi top down door het salesfunnel lopen.

J: Ga er maar vanuit dat het zo is. Volgens mij draait jouw onderzoek om uit te vinden hoe dit proces efficiënter kan worden. Zodat je met minder telefoontjes genoeg bezoeken binnen kunt halen zodat je meer tijd over houdt om het ‘spel onder in de funnel’ te spelen waar het echte geld verdient wordt. Dus ik denk dat het normaal gesproken bellen-->bezoeken-->aanvraag-->plaasten is. Maar dat betekent ook dat we veel waste creeren. We willen naar een optimale funnel toe.

S: Ik kan me voorstellen dat er branches zijn waar mensen bellen leuk vinden, bezoeken ook prima, aanvraag geen probleem. Maar als puntje bij paaltje komt dat er geen geld op tafel komt. Dat is zonde van de tijd.

J: Inderdaad.

S: Op termijn lijkt het mij ook mooi om het model dynamischer te maken om adhv bijvoorbeeld de zij-instroom te kijken naar wat er in bepaalde branches speelt en daar je call lijst op aan te passen.

S: Ja, je hebt heel veel verschillende, bijvoorbeeld relatiebeheer. Deze neem ik niet mee aangezien ik dan verder vertroebel. Zo heb ik een paar weken geleden bij de intercedent gezien dat bijvoorbeeld telefoon-overig de default categorie is voor een contactmoment in mondriaan dus daar vegen ze alles onder als ze even vlug wat invullen.

J: Is er een reden om calls goed te loggen onder acquisitie?

S: Dat zou ik zo niet weten. Maar je kunt eigenlijk kiezen, wat is erger: een false positive of een false negative? Ik denk dat het eerste erger is.

J: waar mijn angst zit is dat relatiebeheer hier te ver buiten beeld raakt terwijl het in mijn ogen een belangrijke rol kan spelen bij potentiële klanten.

S: Ben ik het helemaal mee eens.

J: Ik snap de scope van je onderzoek. In mijn ogen ligt er echter meer potentieel rendement in het doen van goed relatiebeheer en dus het omgaan met bestaande klanten dan in het vinden van nieuwe klanten.

S: Ok.

J: Beetje afhankelijk van de definitie van klant. Bij ons is iemand klant af als hij een jaar geen factuur betaald heeft. Wanneer ga je die bellen?

S: Je zou die een keer per jaar of zo willen bellen?

J: Het liefst per kwartaal ofzo. we hebben onze klanten ingedeeld met een status en een bewerkingscode. Bewerkingscode zegt iets over de omzet potentie. Bijna alle suspects zitten in categorie 4(tot 50k omzet per jaar).

S: Ok.

J: Alles valt of staat bij de juiste registratie.

S: Inderdaad, you measure what you ask for geldt hier heel sterk heb ik het idee.

J: Inderdaad.

S: Bij de intercedent had ik het idee dat mondriaan als een noodzakelijk kwaad gezien wordt.

J: De vraag is of je mensen wilt controleren of wilt helpen.

S: Inderdaad, en of je het verschil kunt duidelijk kunt maken.

J: Ik denk dat als jij een shortlist kunt maken adhv dit onderzoek dat je de intercedenten daar echt mee helpt en dat je ze dan kunt vragen om calls op de juiste manier in te gaan voeren en dat ze het dan ook echt gaan doen. Dan creëer je een win-win situatie. Je moet wel op zoek naar een paar units die hierin mee willen gaan om het te testen.

S: Inderdaad, dan snappen ze het.

J: Het is interessant om te kijken hoe de call data is opgedeeld: of je oververtegenwoordiging ziet in bepaalde branches e.d.?

S: Grappig dat je het zegt, zo is er inderdaad een ondervertegenwoordiging van middelgrote bedrijven. Maar ik denk dat het belangrijker is om een correcte afspiegeling te hebben dan om een volledige afspeling te hebben.

J: Mee eens. Maar het zijn waarschijnlijk wel de sales-tijgers die het doen.

S: Of juist de ‘regel-neukertjes’. Dan hoop je dat het elkaar een beetje uitmiddelt. Dit is een van de limitaties van het onderzoek.

J: Kijk je nog naar unit types?

S: Ik heb al dat soort data. Ben nu bezig met het verkennen van de data. Een ander ding waar ik heel benieuwd naar ben is hoe lang jij zou wachten na een call totdat je het conversie noemt?

J: Het makkelijkste antwoord daarop is om terug te kijken in de data en kijk hoe de curve eruit ziet. Dit kan ik op gevoel niet voorspellen. Ik verwacht wel dat het relatief kort is.

S: Ok, dit ga ik doen. Ander dingetje: welke databronnen verwacht jij dat interessant zijn om een voorspellend model te bouwen. Die zij-instroom lijkt me interessant bijvoorbeeld.
J: Dit lijkt me slimmer om even tegen Jan aan te houden. CBS, ROA(schaarste op de arbeidsmarkt over een periode van 5 jaar) kan interessant zijn. Het issue daarbij is dat de data niet kort-cyclus is en dus geaggregeerd over de cycli. Daar ga je tegenaan lopen. Ik denk dat het heel moeilijk is om externe data te vinden die nauwkeurig genoeg is om toe te passen. BVNL vind ik zelf heel interessant omdat die tijdens een doorsnede van de tijd en het is dus vrij lastig om verschillen te bepalen. Kan bij Cendris opgevraagd worden maar gaat wel geld kosten. Erik van Tempo Team had data over vacatures, Nielsen. Deze data is geloof ik redelijk tijdig.

S: Ik heb zelf ook nog zitten denken aan exotische bronnen als twitter.

J: wat geeft dit voor waarde. Wat voorspelt hij precies, dat je meer moet gaan bellen?

S: Ja, bijvoorbeeld. Hij maakt geen onderscheid tussen branches en bedrijven. Wat hij wel kan doen dat hij groei van de hele economie kan gebruiken om dingen te voorspellen. (hier begin ik onzin te praten waar ik inmiddels verder ben in mijn denken dan vorige week)


S: Ok. Ik heb ook nog gekeken naar leading indicators. Stel dat ik de orderportefeuille van ASML in de tijd zou hebben, die een hele goede leading indicator is voor de economie.

J: Dit zegt iets over de groei van de economie, wat zegt dit nou voor het model?

S: ASML en zijn toeleveranciers in de regio Eindhoven stijgen als ASML zijn portfolio over een jaar goed vol zit.

J: Dit zou dan een leading indicator kunnen zijn voor de geo-as. De vraag of je dit wilt doen op postcode, gemeente, of rayon, arbeidsmarktregeio en welke weging geef je hier aan.

S: Deze zou ik kunnen construeren aan de hand van de longitude en latitude data die ik heb.


Hierna verzandt het verhaal in geouwehoer over hypotheses wat in principe onzin is aangezien dit pas later aan de hand van het model bepaald worden. AF!

---

The Business consultant
S= Sander, A = The business consultant
S: We hebben al veel gesproken dus dit is allemaal meer ter bevestiging.
A: Inderdaad
S: Het interview gaat over twee doelen: call-conversie en externe bronnen. Vooral het laatste deel is hier van belang nog.

A: Ja.
S: Laten we beginnen met call conversie: wanneer is een call nou een succes? Wat ik tot nu toe heb: ik neem alleen de acquisitie calls mee. Een aanvraag neem ik mee tot 21 dagen nadat hij gepleegd is. Wat een ding is dat ik nog toegevoegd heb, is dat er alleen marktunits toegevoegd moeten worden.

A: Dat vind ik een goeie, dit maakt de definitie helemaal stevig. Neem je bestaande klanten mee?
S: Nee, omdat je die niet om alleen acquisitie redenen bezoekt.
A: Ok, mee eens.
S: Dan denk ik dat mijn conversie op orde is.

A: Dan heb je een mooie definitie, scherp en groot genoeg.
S: Ok, dan het tweede stuk. Welke databronnen zijn nou interessant om conversie te voorspellen? Een van de dingen waar ik mee zit is dat bijvoorbeeld verschillen tussen branches lastig te vinden zijn (de voorspellende waarde daarvan).

A: Wat je in feite doet is een momentopname van een tijdsreeks voeden aan je model.
S: Ja, inderdaad, een volledige tijdsreeks kan niet in pure vorm in het model gestopt worden. Ik voeg kolommen toe aan mijn data om het te gaan gebruiken.

A: Je wilt eigenlijk een correctiefactor gaan maken om de verschillende tijdsreeksen op elkaar toe te leggen. Dit zou je kunnen doen mbv een python script.
S: Ok. Wat zijn out of the box bronnen die je zou willen hebben.

A: Het belangrijkste wat er zou zijn alle btw-aangiften van alle bedrijven in Nederland. Dan weten we met lag van een kwartaal hoe het gaat met bedrijven.
S: Dat zou gaaf zijn, maar ik denk niet dat ze die vrijgeven.
A: Wellicht in geaggregeerde vorm.
S: Hoe zou dit ideale bestand benaderd kunnen worden.
A: Merk op dat ik hierbij aanneem dat de groei in een branch gelijk is aan groei van onze omzet binnen die branch.
S: Vertegenwoordig je hierbij niet de flexgraad?
A: Ja, dat is nog een ideaal bestand wat je zou willen hebben. Dit kunnen we wel per branch uit onze systemen trekken. Ik denk dat elke branch zijn eigen flexgraad heeft.
S: Zou de inkoop index ook niet interessant zijn? Die loopt voor op btw-aangiften van bedrijven.
A: De manager inkoop index is heel goed maar van consumenten niet.
S: Ok. Wat ook wel interessant kan zijn is de samenhang tussen branches.
A: Zo’n correlatiematrix kunnen we maken uit onze eigen data.
S: Ok. Heb je verder nog ideale datasets die je zou willen?
A: Ja, in het ideaal geval zou je alle contactpersonen van bedrijven met hun inkoopgedrag willen hebben. Maar dit is vrij lastig te doen. Ook BVNL zou mooi zijn.
S: Ok. Zou ik Cendris kunnen mailen om oude sets van BVNL te kunnen krijgen?
A: Je kunt het proberen. Ook vacaturedata kan interessant zijn. Wat is er per functie per branch en wat wordt er gevraagd per functie per branch. Wat zit er in de kaartenbak van de intercedent.
S: Als het goed is vind ik er exact geen extra conversie mee.
A: Als het goed is ja.
S: Ik heb het zo allemaal wel ongeveer. Bedankt voor je tijd.
A: Graag gedaan.

The manager marketing intelligence
S= Sander, A= the manager marketing intelligence.
S: Wat is jouw functie precies?
A: Manager marketing. Leider van een dedicated team dat voor tempo team werkt. We werken in vier domeinen. Het belangrijkste is dat we continu inzichten houden in alle trends en ontwikkelingen die relevant zijn voor TT en haar opdrachtgevers. Twee keer per jaar brengen we dit uit. Daarnaast hebben we drie andere domeinen. Toegepast onderzoek, daarbij moet je denken aan de publicaties die we maken zoals het rode boekje, Anders Werken. Dat dient als input voor de campagnes. Het derde domein is alles wat te maken heeft met customer experience. Direct feedback systematiek over customer experience. Daarnaast hebben we een model neergezet over ‘personas’ voor de B2C kant (categorieën van klanten). Hier wordt de ‘customer journey’ aan toegevoegd. Dit gaan we ook doen in de B2B kant. Segmentatie is het toverwoord. De laatste categorie is alles wat te maken heeft met data. Daarin ontsluiten we een aantal toolings en datasets en zijn wij op verschillende niveaus actief. Hier moet eigenlijk een nieuw domein bij gaan komen, wat te maken heeft met competitor intelligence.
S: Waar ik mee bezig ben is het maken van een voorspelsmodel. Wat het eigenlijk is is classificatie van B2B bedrijven: welke bedrijven moeten we nou gaan bellen? Binnen de vlinder heb ik gefocust op de outbound calls. Hierbij kijk ik van calls naar visits en calls naar requests.
A: Dit kan wel maar je moet pull units buiten beschouwing laten. Je hebt zeg maar 3 typen units. Een marktaanvraag wordt gekenmerkt door heel veel klanten heel veel aanvragen. Dan heb je account units, met een à twee grote klanten. Een pull unit kenmerkt zich door een paar profielen en meerdere klanten [de horeca]. Deze laatste twee leven onzuivere sales funnels op en zou ik buiten beschouwing laten. Ik denk dat de definitie die je nu hanteert een goede is.
S: Ok, bedankt.
A: [lange verhandeling over de ontwikkeling van de arbeidsmarkt en de uitzendbranche. Heel erg interessant maar niet van belang voor dit onderzoek]
S: En wat betreft externe databronnen.
A: We hebben bij tempo team een grote lijst met bronnen. Misschien kun je daar wat mee.
S: Bedankt. Ik zal deze vertrouwelijk behandelen.

A: Als je wilt gaan voorspellen binnen de uitzendbranche heb je te maken met economische trends, branche gerelateerde trends, factuur statistieken en regio trends. Dat laatste is het geval omdat bedrijven geografische clusters zijn gaan vormen met bedrijven die in hun supplychain zitten.

This interview transcript was not totally complete.

Marketing intelligence analyst
S = Sander, J = Marketing Intelligence analyst
S: Bedankt dat je even tijd kon maken. Zoals ik zei net, waar ik mee bezig ben zijn de definities. Deze is inmiddels zo goed als af. Het tweede stuk gaat over de externe databronnen. Wat is jouw rol precies?
J: Met name die blik van buiten naar binnen brengen. Een paar jaar geleden hebben we de switch gemaakt van inside out naar outside in: wat gebeurt er in de markt en wat benchmarken met de interne prestaties. Wat zijn de trends buiten en haken we daar wel genoeg op in. En dat is enerzijds heel strategisch maar ook heel operationeel. Maar ook PR-matig, richting opinie leider schap. Op dit moment is sales het belangrijkste en daar valt jouw project ook binnen denk ik. Heb rond 2010 veel gefocust op de strategie. Mijn achtergrond is in de psychologie, consumentengedrag. Ik ben nu vooral bezig met BI.
S: Consumentengedrag gaat vooral over user experience of meer op macro niveau.
J: Meer op consumentengedrag. Daar zit een technische component en een ‘softe’ component in.
S: Waar ik naar aan het kijken ben is naar de call conversie. Uiteindelijk gaat het erom om de sales te verbeteren. Ik ga kijken naar de linkerkaart van de vinder. Ik neem bewust de plaatsingen niet mee in mijn conversie. Wat ik nu ga doen met deze definitie is kijken naar de calls naar bedrijven. Ik wil er data aan toe gaan voegen om te gaan voorspellen welke call gaat converteren. Met de interne data vind ik niks. Nu ga ik kijken aan de hand van externe data of ik deze voorspelskracht kan verbeteren. En daar wil ik het met jou over hebben, wat jij denkt dat interessante bronnen zijn. Daar wil ik zo out of the box mogelijk over denken want tegenwoordig is er veel meer data dan we meestal door hebben. Wat is nou echt gaaf om te weten?
J: Wat ik zou willen weten is volgens mij zijn kenmerken van bedrijven. Ik denk dat locatie, bedrijfsgrootte, branche. Een combinatie van deze macro/economische gegevens kan helpen in mijn ogen. De functuur is per branch bekend. Wat is de relatie tussen de macro ontwikkelingen en de calls. High level data is denk ik niet zo interessant. Je wilt veel meer dingen op een gedetailleerd regio niveau weten. Ik denk met name het koppelen van verschillende elementen, industrie aan lokatie, dat nuttig kan zijn.
S: Ik kan dit mbv de postcode kan de bedrijven aan een lokatie ‘mappen’. En de grootte kan ik weer koppelen aan de naam. Uit een van mijn andere interviews heb ik drie assen te horen gekregen die wellicht interessant zijn om alle data op te mappen.
J: Deze assen zijn vooral vrij goed haalbaar. Als je out of the box wil denken zou je bv Richting exotische dingen als nieuws crawl en om bv informatie over uitbreidingen van bedrijven te krijgen.
S: Wellicht dat dit ook gebruikt kan worden voor andere bedrijven in de branch.
J: Het zal vrij lastig zijn om dit te vertalen naar een paar indicatoren. Dat ken ik niet. Maar het lijkt me wel heel interessant. Ik denk veel mensen in hun hoofd wel hebben zitten hoe ‘het werkt’ maar dat is nergens opgeschreven. Wat ik wel heb is een plaatje hoe de plek van branches in de functuur samenhangen. Dat plaatje zal ik even doorsturen.
S: De vraag waar het in mijn ogen om gaat is ‘waarom gaat een bedrijf meer uitzendkracht in het algemeen aannemen en waarom riskeren?’ Wat mijn mijn interessant lijkt is om per branche of supply chain kijken. Over ‘een grote vis’ is veel te vinden, welke bedrijven en branches worden meegetrokken in deze groei.
J: Dat lijkt me interessant. Je hebt wel van die plaatjes over hoe verschillende branches economisch samenhangen. Met deze plaatjes zouden er onlogische dingen uit de dingen die we doen gehaald kunnen worden. Een voorbeeld hiervan is dat er vrij weinig naar groothandel gekeken wordt bij economische groei terwijl dit toch echt wel zo is. Laatst met een lijn-manager gesproken. Die kende bijna geen groothandels. S: Dan zou je ideeën in willen zoomen op een opvallende sector om te kijken wat er aan de hand is.
J: Dit zou ook gebruikt kunnen worden om beter te gaan presteren en dus betere mond op mond reclame te krijgen.
S: Inderdaad. Ik ben ook aan het kijken naar LinkedIn data van intercenten.
J: Zoietst zijn we op het moment ook mee bezig om oude klanten terug te winnen. Oa door te kijken naar LinkedIn. Wat ook interessant kan zijn is het meenemen van klant tevredenheid voor voorspellen.
S: Het nadeel hieraan is dat het om bestaande klanten gaat en ik denk dat dat niet slim is om mee te nemen.
J: Je zou het willen beperken tot niet-klanten?
S: Ja. Ik heb bij de intercenten gezeten en ik merkte dat ze bij het invoeren de weg van de minste weerstand kozen. Dat is echt een limitatie van mijn onderzoek.
J: Je zou ook de omzetpotentie kunnen meenemen wat door Randstad geschat wordt. Het nadeel is dat deze getallen vrij slecht afgeschat worden.
S: Inderdaad. Het lijkt mij zinnig als er uiteindelijk een soort tableau achtige omgeving is waar mensen hun gewenste data uit kunnen toveren.

J: Zo iets bestaat nog niet. Maar zou wel goed binnen het ABFS verhaal passen. Qua externe bronnen zat ik ook nog te denken aan een nog gedetailleerder niveau te kijken naar vacatures uit het JobFeed bestand. Dat is allemaal terugkijken. Maar ik kan me voorstellen dat het in relatie tot S: Dit is meer toepasbaar aan de andere kant van de markt. Wat wellicht ook interessant zou kunnen zijn is de vergrijzing per branche/functie.

J: We hebben het ROA bestand. Daar staat dat soort data ook in. Ik kan je dit wel opsturen. Hier zitten veel kengetallen en prognoses in. Vanuit de Universiteit Maastricht, wij zijn een van de sponsors.

S: Wat zijn nog meer bronnen die je gebruikt?

J: JobFeed, CBS, UWV (is minder relevant voor jou verwacht ik, is meer de rechterkant van de funnel). Daarnaast zijn er een aantal onderzoeken die interessant zijn. Arbeidsmarktverkenning. Een conjuncturenquete van kvk en CBS, daarin vragen ze ieder kwartaal aan ondernemers wat hun verwachting is qua omzet, personeel e.d. Bij banken zijn ook interessante dingen te vinden.

S: Wat ook interessant kan zijn wat het deel van uitzendkrachten per branche in de tijd.

J: Hier heb ik wel een figuurtje van, ga ik je opsturen.