KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

FACULTY OF ECONOMICS AND BUSINESS

Data Mining in Employment

Machine Learning: Predicting Time until Employment

Alexander Derkinderen
r0302563

Carmen Vandeloo
r0301871

Thesis submitted to obtain the degree of

Master in de toegepaste economische wetenschappen:
Handelsingenieur in de beleidsinformatica

Promotor: Prof. Dr. Wilfried Lemahieu
Assistant: Michael Reusens

Academic year: 2015 - 2016

KU LEUVEN
FACULTY OF ECONOMICS AND BUSINESS

FACULTY OF ECONOMICS AND BUSINESS

Data Mining in Employment

Machine Learning: Predicting Time until Employment

Nearly a third of all job seekers have been unemployed for more than two years. A more proactive effort towards significantly reducing such level of long-term unemployment (e.g. through better guidance to job seekers) is essential. And so, the main objective of this thesis is to develop a model that makes it possible to predict upfront who is at risk of long-term unemployment and explore significant attributes that allow for better insights on how to shorten the unemployment duration.

For this purpose, different multiclass classification models have been built based on current data from the VDAB (public employment service of Flanders) by means of a variety of machine learning algorithms used for supervised learning. In general, the performance level of these constructed models is low, with the highest F score at around 0.46. Low-term unemployment proves to be significantly easier to predict than medium-term and to a lesser extent long-term unemployment. In addition, white-box systems perform significantly better than black-box methods in predicting unemployment duration, allowing for more interpretable algorithms to be used. Finally, suggestions for further research include, applying more detailed attributes and more specific models.

Alexander Derkinderen
r0302563

Carmen Vandeloo
r0301871

Thesis submitted to obtain the degree of

Master in de toegepaste economische wetenschappen: Handelsingenieur in de beleidsinformatica

Promotor: Prof. Dr. Wilfried Lemahieu
Assistant: Michael Reusens

Academic year: 2015 - 2016

Acknowledgements

Leuven, 11/05/2016.

This research was supported by the public employment service of Flanders (VDAB) which provided us with much relevant data that was key to addressing the problem at hand. Moreover, useful insights into the specific business context of unemployment were also shared with us. As such, we would like to thank them for their expertise and support, in particular Erik Klewais, Stijn Van De Velde, Michael De Blauwe, Ellen Van Molle, Joost Bollens and Martine Geers. At the same time, IBM was also conducting research on the issue of unemployment duration, and so we are also grateful to Kristien Verreydt, Jan De Beer and Max Carakehian for taking the time to discuss some key issues related to this research.

We would like to express our deepest gratitude to Michael Reusens, supervisor and PhD student at the Faculty of Economics and Business, for always standing by to guide us through any difficulties (R programming...) and to follow up on our progress. His feedback helped us delve deeper into certain topics and made us more aware of motivational decision making, for example. We hope this research can in some way contribute to his final dissertation. Beside him, Professor and Vice-Dean for Education at the Faculty of Economics and Business, Wilfried Lemahieu, charted for us new paths to discover en route to finalizing this research. And so, we would also like to take this opportunity to thank him for his feedback on our research proposal, related work and final draft.

Finally, we would like to thank all those who lent a helping hand with the proofreading of this thesis and those who provided us their love and support.

Table of Contents

Acknowledgements **i**

Management Summary **1**

1 Introduction **2**

2 Research Questions **4**

3 Related Work **5**

4 Research Setup **8**
4.1 Data Selection & Preparation 8
4.2 Data Transformation 9
4.2.1 Case 1 9
4.2.2 Case 2 10
4.2.3 Case 3 10
4.2.4 Case 4 10
4.3 Data Mining Techniques 11
4.3.1 Overview of Machine Learning Techniques 11
4.3.2 Overview of Samples 14
4.3.3 Evaluation Metrics 14

5 Results **16**
5.1 General Samples 16
5.1.1 Cox PH Model 16
5.1.2 Predictive Models 17
5.2 Subsamples with Temporary Work 18
5.3 Subsamples without Temporary Work 19

6 Benchmarking **21**
6.1 Comparing Data Mining Techniques 21
6.2 Comparing Time Intervals 24

7 Recommendations to the VDAB **26**

8 Conclusion **28**

Appendices **29**

List of Figures **45**

List of Tables **47**

Bibliography **48**
Articles 48
Books 51
Internet 52

Management Summary

The Public Employment Service of Flanders - also known as VDAB - is tasked with guiding job seekers in their search for employment [1]. To be more specific, a consultant of the VDAB assists people in their search for a job based on his/her knowledge accumulated throughout his/her career. However, information dating from December 2015 [2] reveals that nearly a third of job seekers have been unemployed for at least 2 years. Given this situation, a proactive person-specific approach could influence the guiding process in a more positive manner than the mere expertise of a VDAB consultant.

Unemployment data was obtained from the VDAB in order to build multiclass classification models. These models seek to ascertain patterns in the data that allow for the newly unemployed to be correctly predicted in terms of time intervals in unemployment. Time intervals were acquired from a business context and set at Low (0-3 months), Medium (3-6 months), and High (>6 months). In this way, people who are at risk of long-term unemployment could be identified early in the process.

To build those models, the 'raw data' had to undergo some form of transformation and reduction. This resulted in a final dataset of 33 attributes and 1,352,446 instances. Afterwards, eleven prediction methods were trained on a randomly sampled trainset to predict at registration the unemployment duration of a job seeker based on his/her distinctive personal characteristics. In addition, most of the trained models were able to provide insights on why a certain job seeker would be at risk, e.g. lowly educated. These kinds of insights are derived from white-box algorithms, and can be expressed in an intuitive manner, e.g. rules or decision trees. These outcomes can then be used by the VDAB to provide guidance to job seekers in a more proactive manner.

However, due to a low predictive performance by the initial models, others were trained using different subsamples. That is, the data was subsampled by applying filters based on age and education. More precisely, the data was split into different age categories: younger than 25, between 25 and 50, and older than 50; as well as different educational levels: low, medium, and high. This, however, did not have any significant impact on the predictive performance measure - macro-averaged F score - of the models.

All models were evaluated based on overall performance through their macro-averaged F score, which is a combination of precision and recall. On the whole, the trained models' performances are rather low, with the highest scores at around 0.46 and the highest possible value at 1. However, short-term unemployment and to a lesser extent long-term unemployment are easier to predict than medium-term unemployment. It is thus possible to provide job seekers at risk of long-term unemployment with more effective guidance, as these are to a certain extent distinguishable from those with short-term unemployment. One ought to exercise caution though when drawing conclusions based on the constructed models, as so far all without exception have been underperforming.

In conclusion, further research paths may include increasing the number of filters in order to train more specific and smaller models, adding more detailed information on certain attributes that have been aggregated in this thesis (e.g. degree of education), and incorporating attributes that describe job seekers' social and soft skills, motivation, and willingness to work.

Chapter 1

Introduction

The public employment service of Flanders (also known as VDAB) was founded in 1989. Inherent to its foundation is the need to bring together supply and demand in the job market. Hence, its main task is to assist job seekers in their search for employment [1]. And so, job seekers are guided through dialogue with a consultant. This consultant is an employee of the VDAB, whose task is to guide people in their search for a job based on his/her past experiences. This, however, is a rather general approach that makes it hard to account for all the relevant characteristics of a job seeker.

At the end of 2015, there were 228,987 non-working job seekers (NWWZ) in Flanders according to data from the VDAB [2]. Table 1.1 shows how certain categories based on age, education level and unemployment duration are represented among these non-working job seekers. According to the Belgian Ministry of Economic Affairs, 5.2% of the population in Flanders were unemployed in 2015. More detailed information can be found on their website [3].

		dec/15	year difference	share
NWWZ		228,987	-0.30%	
Age	- 25 y	44,959	-0.30%	19.60%
	25 - 50 y	120,611	+3.00%	52.70%
	+ 50 y	63,417	+5.20%	27.70%
Education	Low	105,397	+0.80%	46.00%
	Medium	83,507	-0.40%	36.50%
	High	40,083	-3.00%	17.50%
Unemployment duration	- 1 y	121,480	-1.90%	53.10%
	1 - 2 y	39,163	-7.80%	17.10%
	+ 2 y	68,344	+7.80%	29.80%

Table 1.1: Unemployment data by VDAB (December 2015)

Within the age categories, two relatively large groups emerge: - 25 years old and + 50 years old, representing respectively 19.60% and 27.70% of the 228,987 job seekers. Taking into account that the minimum legal working age is 15 for part-time work and 18 for full-time work, this means that the -25 years old category consists of a much smaller age range than the 25-50 years old category. As a result, a share of 19.60% is quite high. The same reasoning, combined with a rising share of 5.20%, makes the +50 years category the second of the two aforesaid large groups. Note that the legal retirement age in Belgium is set at 65 [4]. Moreover, it becomes more difficult to find a job beyond the age of 50 [5]. Consequently, both groups will be highlighted in Ch. 5, where results are discussed.

The main issue that presents itself here is that nearly a third - and still rising - of all non-working job seekers have been unemployed for more than two years, as shown in Table 1.1. In general, the longer a person is unemployed, the harder it is to find a job [6]. Therefore, it is crucial to provide better guidance to job seekers and to detect more quickly those that pose a risk in terms of long-term unemployment. In addition, a more person-specific approach is needed to function as guide instead of solely relying on the general expertise of a VDAB consultant.

The **problem description** is structured in the following part. When people look for a job they register themselves - sometimes mandatory - at the VDAB. From the moment a job seeker registers as unemployed, we would like to know is whether s/he poses a risk in terms of long-term unemployment. A number of analyses have already been conducted on this, each with emphasis on asserting why a certain person is unemployed for a longer period than another [7][8][9]. In this research, we mainly focus on predicting the time interval between registration and the taking up of a job. Hence, the problem can be structured as a survival analysis. Survival analysis encompasses different methods which analyse time in terms of event occurrences, as illustrated in Figure 1.1. Here, the outcome attribute is the time needed by a specific person to find a job and the event is 'job found'. One can also think of other domains where problems can be structured as time to event (see infra, Ch. 3).

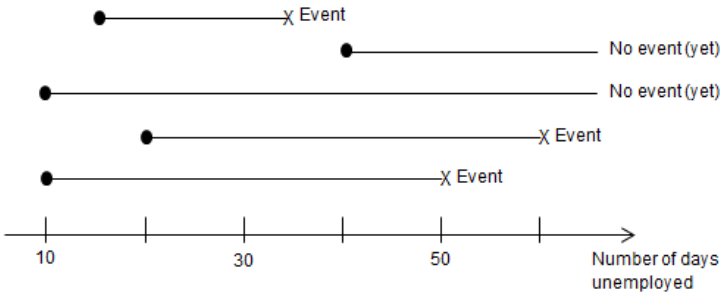

Figure 1.1: Time to event instances

In the context of machine learning, the problem of predicting unemployment duration among new job seekers can be modelled as a supervised classification task. In supervised learning, patterns are derived from a dataset in order to classify instances (unemployed people) in their appropriate class [10]. The algorithms keep on learning - and improving - by adding new instances to the dataset. In this way, models, based on current data, try to give an accurate indication of how long a new job seeker will be unemployed, given his/her characteristics. In addition, models should be somewhat interpretable to VDAB consultants, in a way that enables them to provide clear explanation to job seekers about the risk of ending up in long-term unemployment. A small hypothetical example will try to clarify the above.

> *Alexander recently graduated from university and enrols at the VDAB as a job seeker. He has a driving license B, Dutch as mother tongue, and is 22 years old. There are of course many other characteristics that can be taken into account such as residence, knowledge of French, etc. The VDAB registers all of Alexander's details, puts them into (different) models, and as a result obtains the predicted unemployment period. Preferably an indication is given as to why this certain period is predicted, e.g. time period is low because he has a driver's license. Based on this duration and a number of business rules, the VDAB is able to determine in an objective way whether or not to (intensively) guide Alexander in his search for a job.*

This proactive approach should have an improving effect on actions taken by the VDAB, which are in a sense beneficial to both job seeker and employer. For example, the model could suggest that taking up a certain course of training would lower the unemployment duration. The VDAB could then recommend that specific course of training to the job seeker, from which an eventual employer will one day benefit. The VDAB can also take advantage of this approach by ensuring a more efficient and effective allocation of resources.

Chapter 2

Research Questions

In general, we would like an answer to the following question: How long does it take for a specific person to find a job after s/he becomes unemployed at a certain point in time? Moreover, we are also interested in finding out why it takes 'that amount of time' to find a job. Both questions can be combined and made more specific in order to define a clearer scope to this study. More precisely, the main question of this research is formulated as follows:

> *Which data mining techniques can be used to predict and explain unemployment duration among the newly unemployed?*

The latter is split up into four sub-questions, which are used as a guide throughout the process of ascertaining results; these are also aligned with the different steps of knowledge discovery in data, as explained in the research setup (see infra, Ch. 4).

The first sub-question addresses the importance of attributes in the dataset: **which attributes have an impact on the performance of predictive models?** Feature or attribute selection is a very important step in reducing the dimensionality of the dataset in order to build simpler and more predictive models [11]. In this research, many attributes were provided in the dataset. However, not all of them were useful for predicting unemployment duration or had any impact on the models' performance. This first sub-question is discussed in chapter 4, the data selection and pre-processing step.

The second sub-question refers to the performance of different techniques that were applied: **which machine learning techniques have a high performance in predicting time until employment?** In total, 11 algorithms were chosen based on their proven track record, as discussed in chapter 3 Related Work. A variety of techniques were considered to address this sub-question: decision trees, neural networks, boosting, support vector machines, etc. All of them are described in section 4.1. Note that white box models (DT, Rule learners, etc.) are preferred over black box models (NN, SVM, etc.) because a specific result should be traceable. Therefore, more white-box algorithms are considered.

Thirdly, exploring is done over the main reasons why someone would take a certain amount of time until employment: **which attributes are significant in explaining the predicted unemployment duration?** Cox proportional hazard model is performed to check for any significant attributes. It is essential that the VDAB is able to provide a very clear explanation over the outcomes of this model. In this way, they would be able to cope with difficulties one might have at the beginning of his/her unemployment, e.g. an inadequate understanding of Dutch, a rather low level of education, etc. These results are briefly discussed in chapter 5.

Last but not least, techniques are compared and trade-offs are made: **how do the different data mining techniques relate to each other?** This will allow for the selection of the best algorithm, given the different methods that are applied. Clearly, a trade-off should be made between performance on the one hand and interpretability or comprehensibility of the model on the other. This trade-off is discussed in chapter 6.

Chapter 3

Related Work

Survival analysis is used in multiple areas, as indicated by the following diverse research topics. J. Lu (2002) [12] focused on customer churn rate in the telecommunications industry. More precisely, customer survival functions were estimated using the accelerated failure time (AFT) model in SAS. Binary class labels were used to predict whether someone churned or not. Similar research was conducted by J. Lu & O. Park (2003) [13] and more recently by L. Fu & H. Wang (2014) [14]. The latter explored the Cox proportional hazard model in an insurance-based setting. Other areas can be envisaged as well: engineering (e.g. time until failure of some component) or social behaviour (e.g. time until couples divorce).

However, most of the applicable techniques are discussed in research about medical prognosis. Survival analysis is widely used to predict the survivability chances of diseases, e.g. cancer. A first strand of literature includes W.C. Levy et al. (2006) [15], L. Franco, J.M. Jerez & E. Alba (2005) [16], J. Llobera et al. (2000) [17], A. Barth, L.A. Wanek & D.L. Morton (1995) [18], P.C. Adams, M. Speechley & A.E. Kertesz (1991) [19], C. Rozman et al. (1984) [20], and J.L. Binet et al. (1981) [21]. These cover the use of the Kaplan-Meier (KM) estimator to estimate survival functions from datasets containing patient records. In addition, Cox proportional hazard model has been used to test the significance of different variables related to time until event.

In recent decades, a second strand has discussed the use of machine learning algorithms to predict survival times. S.S. Anand et al. (1999) [22] have compared an artificial neural network (ANN), regression tree, and a manipulated K-nearest neighbour algorithm with the Cox regression model. Despite the fact that ANNs can model non-linear relationships, results indicated that Cox regression still performs best when a dataset contains censored instances. This immediately conjures up one of the major problems with machine learning in a survival-based setting: censored data. Different approaches have been used to overcome this hurdle. L. Bottaci et al. (1997) [23] have simply ignored censored instances, while others have assigned them to separated groups learning as a true/false classifier; H.B. Burke (1994) [24], H.B. Burk et al. (1997) [25], L. Ohno-Machado (1997) [26], and A. Bellaachia et al (2006) [27]. However, these methods do not address the issue of censored records directly, and may involve information loss in particular settings, e.g. small datasets.

M.D. Laurentiis & P.M. Ravdin (1994) [28] have suggested a framework for incorporating censored cases. More specifically, they proposed a transformation of the dataset into different time intervals and then use KM to estimate the survival probabilities for each of these intervals. They applied their approach to neural networks and found that "NN work well in producing predictive models in situations where Cox regression has some limitations". B. Zupan et al (2000) [29] used this framework when comparing the Naive Bayes classifier and decision tree with the Cox regression model. Results indicated that Naive Bayes performs equally to the conventional Cox PH model in a prostate cancer dataset. An important limitation of the suggested framework is the inability to predict survival times for individual instances. N. Street (1998) [30] addressed this issue by formulating an alternative model that integrates censored cases "directly into the training set" and does not use artificial time intervals. Street's approach has been applied by C. Chi et al. (2007) [31]. Other modifications of ANNs so as to manage censored data have been listed in B. Baesens et al. (2004) [32]. Although neural networks are widely discussed in the literature, they are not very useful for this research because of their "black box nature" as

described by J.V. Tu (1996) [33]. Worth remembering is that acceptance of our model by the VDAB largely depends on the comprehensibility of the results.

Other interpretable data mining techniques include rule- and tree-based classifiers, logistic regression models, and to a lesser extent ensemble methods. The latter have the advantage of performing better in terms of accuracy than single decision trees, as described by A.T. Azar & S.M. Metwally (2012) [34]. G. Llczuk & A. Wakulicz-Deja (2005) [35] discussed different kinds of rule- and tree-based algorithms for a medical diagnosis system: Ridor, J48, PART, JRip, etc. These types of classifiers were chosen for their proven track record in medical decision systems and their interpretability for humans. As a result, new patients can be classified based on intuitive rules.

When higher predictive performance is necessary, one can apply the ensemble methods as described by A.M. Prasad, L.R. Iverson & A. Liaw (2006) [36] and by B.P. Roe et al. (2005) [37] Bagging, Boosting, Voting, and Stacking. A.T. Azar & S.M. Metwally (2012) [34] have applied three types of decision tree classifiers to a breast cancer dataset: Single decision tree (SDT), boosted decision tree (BDT), and random forest (RF) which is a variation on bagging. BDT scored better on all performance criteria - accuracy, recall and ROC - than SDT. However, SDT is a more comprehensive technique and easier to visualize. Clearly, a trade-off should be made here. Y.H. Chan (2005) [38] has given an overview of multinomial logistic regression models (MLR). These models are capable of handling dependent variables with more than two levels. MLR is able to predict the outcome for new instances based on categorical and continuous variables, and does not adopt many assumptions according to J.A. Anderson (1982) [39]. B. Kempen et al. (2009) [40] used multinomial logistic regression to update soil maps based on previous measures. Another application is discussed by Y. Wang (2005) [41], who evaluates an MLR model for anomaly intrusion detection.

Other techniques used in literature to classify instances are the Support Vector Machines (SVM) and Naive Bayes. The former is mathematically more difficult and results are less interpretable than those of the latter. B.K. Bhardwaj & S. Pal (2011) [42] applied Naive Bayes to predict students' academic performance based on different variables. They split up the response variable into five intervals so as to be able to classify the scores. On the other hand, SVM have been used in a credit rating analysis by Z. Huang et al. (2004) [43]. More precisely, they classified companies according to their credit rating into different groups and found that SVM outperform logistic regression and even have a slightly higher accuracy than ANNs.

In this study, different statistical (e.g. Cox PH model) and machine learning techniques (e.g. DT, NN, etc.) are applied to an employment dataset in order to build a model that can classify the newly unemployed according to different time intervals. One possible outcome is that the VDAB can use these results to allocate resources more efficiently. Similar research has already been conducted by V. Ciuca & M. Matei (2010) [7] in Romania. However, their study was mainly focused on finding the explanatory variables for predicting time until employment, while this research is more focused on predicting this 'time'. Overall, they found that age and education seem to have the most predictive power. Unlike our research, their dataset was very limited in terms of dimensionality and they only used a single technique: Cox regression.

Finally, an overview is made in Table 3.1. Different aspects and techniques are related by means of the several research methods mentioned above. Occasionally, a specific reason for using a certain technique is mentioned between brackets.

Aspects \ Techniques	Decision/Rule Tree	ANN	SVM	Ensemble(b&b)[a]	RF	NB	Logistic regression	Cox regression	KNN
Binary vs Multiclass	Binary[27][34] [36][29][35]	Binary[31][27] [23], Multiclass[32] [26][43]	Multiclass[43]	Binary[34]		Binary [27][29], Multi-class[42]	Multiclass[41] [38]	Multiclass [32][26]	
Censored	[[29], Kaplan-Meier estimates]	[[22], addresses censored data] [[31], Kaplan-Meier estimates]				[[29], Kaplan-Meier estimates]		[[22], addresses censored data][[12], addresses censored data]	[[22], addresses censored data]
Performance measure	Accuracy[27] [34][36][29] [35], Precision [27], Recall [27], Kappa[36], F-score[34], ROC[34]	Accuracy[22][27][32] [23][43], Recall[27] [26], Precision[27] [26],ROC[16], Wilcoxon Test [31]	Accuracy[43]	Kappa[36], Accuracy[34], F-score[34], ROC[34]	Kappa[36]	Accuracy [27][29], Recall [27], Precision [27]	ROC[41], Accuracy [38]	Recall[26][17], Accuracy[32][22][17], Precision [17], ROC[16]	
Context (Medical, Banking, Corporate, etc.)	Medical [[29], simplicity & acceptance][[27], good performance][34][35], Geographical [36]	Medical [[31] , non-linear relationship][[22], addresses censored data][[27], good performance][26][16], covariates & non-linear][23], Banking [32], Corporate[43]	Corporate[43]	Medical[34], Geographical[36]	Geographical [36]	Medical [[29], simplicity & acceptance][[27], fast], Education[[42], easy to use]	Anomaly detection [41], Medical[38]	Unemployment [7], Medical [22][17][21][26] [16], Banking[32][14], Telecom[12][13]	Medical [22]
Prediction vs Explanatory	Prediction [22][27][29][36] [34][36][35]	Prediction[[31], non-linear relationship][22][27][32][16][43] [23] [26]	Prediction[43]	Prediction[36][34]	Prediction[36]	Prediction [27][29], Explanatory [42]	Prediction[41] [38]	Explanatory [7][21][14], Prediction [17][32][26][16] [12][13]	Prediction [22]
Regression vs Classification	Regression [[22], survival time][36], Classification[34][27][36] [29][35]	Regression [[22], survival time][16], Classification[27][43][23][31][32] [26]	Classification [43]	Regression[36], Classification[34]	Regression[36]	Classification [27][29]	Classification [41][38]	Classification [32][26],Regression[22][16] [12][13][17]	Regression [22]

Table 3.1: Overview of related works

[a] b&b: different boosting and bagging algorithms (except random forest).

Chapter 4

Research Setup

The research setup is structured according to the knowledge discovery in database (KDD) process described by Fayyad et al (1996) [44]. This process consists of different steps: data selection, pre-processing, transformation, mining, and finally, interpretation of the results. Afterwards, it is possible to adapt different steps - as illustrated by the grey dotted line - based on feedback from new findings. Figure 4.1 shows these different steps and the relation between them. The goal is to extract new knowledge from the input data and use this knowledge to create value, e.g. efficiently allocating resources to the newly unemployed, predicting the occurrence of cancer in order to start treatment very early and thus raise the survivability chances, etc. The KDD process

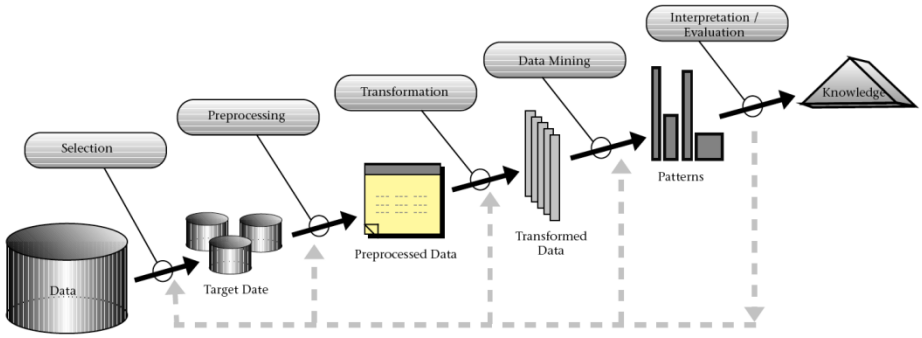

Figure 4.1: KDD process by Fayyad et al (1996)[44]

4.1 Data Selection & Preparation

Data was gathered from the VDAB. More precisely, a spreadsheet was obtained containing 102 attributes and 2,064,192 records. Each record represents a person who enrolled at the VDAB. Multiple records can concern one person as it is possible that after finding a job, one can again become unemployed. The data spans from the beginning of 2012 until the end of 2015. Note that when the data was extracted, some people were still unemployed. As a consequence, some records do not have a value for *catwz_uit*, meaning that they have yet to be signed out at the VDAB. This phenomenon is called censored data, which was referred to in chapter 3.

As our research question is about predicting the unemployment duration of new entrants (into unemployment), attributes that do not have a value at the beginning of this period are not useful and should be removed from the dataset. Further reduction in dimensionality includes attributes that have a constant value, are redundant or highly correlated as well as attributes that can be aggregated. Due to a malfunction of some algorithms when applied to the data with missing values, some numerical attributes have been transformed into true/false values. All attributes are tabulated in appendix 1. The second column indicates whether an attribute is retained from the dataset, and the underlying reason is described in the third column. After having reduced the dimensionality of the dataset into 33 attributes - including the class label -, some transformations were made to the records.

4.2 Data Transformation

Each record in the dataset contains three key attributes which uniquely identify someones unemployment duration: *klnr*, *catwz_in*, and *catwz_uit*. *Klnr* is simply an ID number to represent a specific person in the dataset. *Catwz_in* contains a code that indicates the way in which someone signed up at the VDAB, e.g. code 3 means that a person signed up on a voluntary basis. Similarly, *catwz_uit* represents codes for leaving the VDAB, e.g. code 79 means that a person signed out at the VDAB without finding a job. A list of all codes can be found in appendix 2. A simplified example of a single record is shown in Table 4.1.

KLNR	Date_in	Date_out	Catwz_in	Catwz_uit	#_Days_Unemployed
42561	1/01/2013	25/03/2013	3	79	83

Table 4.1: Example of a record

However, in this research we are only interested in examining the unemployment duration of people who were really unemployed but eventually found a job or are still looking for one. In other words, not all codes from *catwz_uit* are reliable proxies to represent 'job found'. The same holds for *catwz_in*, wherein some people are enrolled at the VDAB not because they do not already have a job but because they want to make a switch. This translates into redefining the unemployment period for certain people. More precisely, four distinct cases related to *catwz_uit* are recognized by the IBM team at VDAB in which data has to be transformed, each of which is explained separately below. In order to cope with *catwz_in*, all people who already had a job when signing in at the VDAB were filtered out of the dataset. Before the data was transformed, 220 observations were removed because there the age was above 65, which is the legal age of retirement.

4.2.1 Case 1

Every record that does not have good successors is checked. Note that good successors are records with the same *klnr* and have appropriate proxies for 'job found'. If those records contain inappropriate proxies, they are removed from the dataset. Inappropriate proxies are defined from a business context and do not represent 'job found'. More specifically, a list of inappropriate values for *catwz_uit* can be found in appendix 3. For example, code 77 means that a person has not found a job but signs out because s/he wants to continue with studies. Note that this process has to be executed iteratively until no more changes occur in the dataset. Table 4.2 gives one example of records that will be removed because they have inappropriate proxies and no good successors (**bold** format of *catwz_uit*). Currently, interim work is regarded as 'job found'. As explained further in section 4.3.2, one can argue that a person who has found a job only for a day is not actually employed.

KLNR	Date_in	Date_out	Catwz_in	Catwz_uit	#_Days_Unemployed
5	1/01/2015	23/02/2015	3	**77**	53
5	1/03/2015	23/04/2015	3	**77**	53
6	1/05/2015	28/06/2015	3	78	58

Table 4.2: Removals in case 1

4.2.2 Case 2

After applying an algorithm to cope with records of case 1, it is still possible for inappropriate proxies to occur. Tables 4.3 and 4.4 provide examples of this behaviour. Remaining records with wrong proxies can be split into two groups, of which the second is discussed in case 3. The first group consists of people who returned to unemployment without any fundamental change(s) in their characteristics. This is often the result of a mere database manipulation, e.g. to get a different *catwz_in* for a person. Consequently, their unemployment duration is added to the next period. Note that the timespan between two periods is not accounted for and could result in an underestimation of the total unemployment duration. This timespan is not included because the probability of finding a job in that period was nihil.

KLNR	Date_in	Date_out	Catwz_in	Catwz_uit	#_Days_Unemployed
4	6/01/2014	14/02/2014	14	**79**	39
4	14/02/2014	4/12/2015	0	78	689

Table 4.3: Removals in case 2

4.2.3 Case 3

The second group consists of people who returned with changed characteristics, e.g. dropped out of the VDAB to continue studying. When they return to the VDAB, a new record is registered and one unemployment period is added. The previous record is removed, as illustrated in Table 4.4. Codes included are listed in appendix 3. Again, one may argue that this transformation could underestimate the actual unemployment duration.

KLNR	Date_in	Date_out	Catwz_in	Catwz_uit	#_Days_Unemployed
4	6/01/2014	14/02/2014	14	**77**	39
4	14/02/2014	4/12/2015	0	78	689

Table 4.4: Removals in case 3

4.2.4 Case 4

Finally, censored records are removed if their duration is lower than the lower boundary of the highest class label (see Table 4.5). This boundary is extracted from the business context in which this research is posited. We can justify this approach thanks to the large dataset that remains after all the removals and the fact that the distribution of this dataset is nearly the same as in the original population. Apparently, most censored cases were already unemployed for more than six months. Note that the inconvenience of censored data is resolved in a context specific way.

These different data transformations reduced the total number of records from 2,064,192 to 1,352,446. As described in Table 4.5, three class labels were used to predict new unemployment durations instead of the actual number of days. The use of discrete intervals is necessary in a supervised classification task.

	Class label		
	Low	Medium	High
Interval	0-3 months	3-6 months	>6 months
Distribution	58.34%	16.42%	25.24%

Table 4.5: Class labels

As mentioned above, some algorithms are not capable of handling missing values. They overcome this hurdle by only taking into account records that have a value for each attribute. In order to perform a useful benchmarking exercise, missing values were replaced by NaNa or -1, depending on the attribute type. In this way, all algorithms were trained and tested with the same records. However, to speed up the training process, a random sample was drawn from the 1,352,446 records, resulting in a final sample of 135,245 records. Note that this threshold of 10% was chosen arbitrarily.

4.3 Data Mining Techniques

We have applied a 5-fold cross validation to the training set. One may argue for applying more than five but as we have quite a large sample, splitting up the dataset into ten parts would greatly increase the computing time. Moreover, some tests with 10-fold cross validation did not lead to significant improvements. Whenever possible, parameters are optimized using the caret package in RStudio®. More precisely, different combinations of tuning parameters are used to train the models. The final model is chosen based on Cohen's kappa statistic rather than accuracy because of the imbalanced dataset [45]. Finally, a test set, which accounts for 30% of the final sample, is used to evaluate the best model's performance. If techniques have only one tuning parameter or even none, then the underlying packages are directly addressed rather than through caret.

4.3.1 Overview of Machine Learning Techniques

Each machine learning technique for building classification models in this research is briefly discussed.

Decision trees are a supervised learning model that hierarchically maps a data domain onto a response set. It divides a data domain (node) recursively into two subdomains such that the subdomains have a higher information gain than the split node [46]. **J48** is an open-source Java implementation of the C4.5 algorithm in the Weka data mining tool. C4.5 is a program that creates a decision tree based on a set of labelled input data [47]. The J48 algorithm was used from the RWeka package. **C5.0** is an update of the C4.5 algorithm. It includes all functionalities of C4.5 and applies new technologies, the most important among them being boosting [48]. The C5.0 train method of the caret package was used, and two parameters were tuned: the number of boosting operations and feature selection [49]. **Rpart** implements many of the ideas found in caret, and Rpart programs build classification and regression models of a very general structure using a two-stage procedure; the resulting models can be represented as binary trees [50]. The 'rpart' train method of the caret package was used, tuning the complexity parameter [49].

Bagging , also known as bootstrap aggregating, is an ensemble machine learning method for generating multiple versions of a classifier and aggregating these different predictions into one. In terms of classification, the majority vote will be the predicted class label. "The

multiple classifiers are formed by making bootstrap replicates of the training set and using these as new learning sets" according to Breiman (1994) [51]. The 'bagging' method was used from the iPred package, incorporating recursive partitioning (Rpart) as tree algorithm [36]. In order to save computing time, the number of cross folds was set to zero and the number of bootstrap replications was held constant at 25.

Random Forest(RF) is an ensemble of unpruned classification or regression trees created by using bootstrap samples of the training data and random feature selection in tree induction. Prediction is made by aggregating (majority vote or averaging) the predictions of the ensemble [52]. The Random Forest model was trained using the 'randomForest' package in R; it implements Breiman's random forest algorithm based on Breiman and Cutlers original Fortran code [53]. The number of variables randomly sampled as candidates at each split was set to three, and 300 trees were grown for each model. According to Oshiro, Perez & Baranauskas (2012) [54], significant performance increases stop at 64 - 128 trees. However, their largest dataset had around 3,500 instances, so we tripled the average number of trees because most of our samples have around 10,000 records.

Gradient Boosting constructs a single model based on multiple base learners. More precisely, different weak learners (trees) are combined into a single strong learner. The difference between boosting and bagging is that the latter takes bootstrap samples and trains learners on each sample, whereas the former uses the entire dataset to train different learners. Misclassified instances are here given more weight so that the next learner would try to classify them correctly [55]. The 'gbm' method from the caret package was used to build the classifier. Four parameters were tuned: number of boosting iterations, maximal tree depth, learning rate or shrinkage, and minimum terminal node size [56].

JRip is a propositional rule learner based on incremental reduced error pruning (irep). It was described by Cohen (1995) [57] as an optimization of irep: a fast effective rule induction learner. The method is more known under the name of repeated incremental pruning to produce error reduction (RIPPER). As in irep, a rule is pruned immediately after composition and then added to the rule set. However, some modifications were made for RIPPER, e.g. ordering classes as well as replacement and revision rules [58]. JRip - to our knowledge - has only been implemented in the Weka tool of Waikato University. Fortunately, the RWeka package provided an interface between R and Weka. As such, the JRip method was addressed directly. Two parameters were set: the number of folds for reduced error pruning which was set to equal the number of cross folds (= 5), and the minimal weights of instances within a split, which equalled two.

Single hidden-layer neural networks are feed-forward neural networks with one hidden layer, which can consist of different hidden nodes. In general, the more nodes and layers, the more complex are the concepts a perceptron can learn. Multi-layer perceptrons are also available in R packages but much more time-consuming to build. Moreover, single hidden layers were used more often in related work. Training is performed by the Broyden-Fletcher-Goldfarb-Shanno algorithm [59][60]. More thorough information regarding neural networks is provided by Huang, Chen & Babri (2000) [61]. The 'nnet' method from the caret package was used to build a neural network. There were two parameters to optimize: the number of hidden nodes present in the hidden layer, and the weight decay [62] [63]. Note that before training, the input data was normalized to overcome biased attributes [64].

Support Vector Machines (SVMs) are a set of related methods for supervised learning, applicable to both classification and regression problems. An SVM classifier creates a maximum-

margin hyperplane that lies in a transformed input space and splits the example classes while maximizing the distance to the nearest cleanly split examples. The parameters of the solution hyperplane are derived from a quadratic programming optimization problem [65]. When no linear separation is possible, a non-linear mapping into a higher dimensional feature space is realized. The hyperplane found in the feature space corresponds to a non-linear decision boundary in the input space [66]. The 'svmRadial' model from the train function in the caret package was used to build models based on support vector machines with a radial basis function (RBF) kernel. This type of kernel was chosen based on research by Huang et al (2004) [43]. There were two parameters to optimize: sigma and cost [49].

Naive Bayes classifier is a machine learning technique based on Bayes rule and a set of conditional dependences, as described by Mitchell & Hill (2015) [67]. More information on the Bayes rule (theorem), together with a clear example, is provided by Triola (2010) [68]. In order to compute the different probabilities, the 'nb' method from the caret package was used. Two parameters had to be tuned: Laplace correction and the type of distribution. More precisely, Laplace correction dealt with zero probabilities while distribution type was set to the Gaussian distribution [69].

Multinomial logit models are an extension of the binary logistic regression (LR) models. Unlike binary LR models, multinomial logit models are capable of handling multiclass response variables. Moreover, the latter contains a variety of models, including the cumulative logit model or proportional odds logistic regression model (OLR). It relates an ordered multiclass response variable to predictors [70] [71]. The 'polr' function from the MASS package in R is used to train such a cumulative logit model. More technical details can be found in CRAN [72]. Note that there are no tuning parameters to optimize.

4.3.2 Overview of Samples

First, all techniques were applied to the first sample: *All*. Afterwards, new samples were drawn based on the most explanatory attributes found by V. Ciuca & M. Matei (2010) [7], namely age and education. Table 4.6 presents the different samples used. The different thresholds for age are derived from the business context. Samples were drawn randomly from the entire population - limited by age and education - with some preservation of class distributions through the 'createDataPartition' function in the caret package [73]. Sample size was set at about 10% of the *All* sample (~10,000) to speed up computing time.

Sample	Age	Education	# Records	Controls
All (- 89)	All	-	135,245 (86,002)	5-fold
All_Laag (- 89)	All	Low	10,951 (11,912)	5-fold
All_Midden (- 89)	All	Medium	11,215 (11,976)	5-fold
All_Hoog (- 89)	All	High	11,592 (10,060)	5-fold
J_25 (- 89)	<26	-	10,110 (11,380)	5-fold
J_25_Laag (- 89)	<26	Low	11,505 (10,801)	5-fold
J_25_Midden (- 89)	<26	Medium	11,377 (10,719)	5-fold
J_25_Hoog (- 89)	<26	High	11,306 (11,119)	5-fold
J_50 (- 89)	>49	-	10,004 (11,157)	5-fold
J_50_Laag (- 89)	>49	Low	10,477 (10,464)	5-fold
J_50_Midden (- 89)	>49	Medium	10,444 (11,435)	5-fold
J_50_Hoog (- 89)	>49	High	10,690 (10,925)	5-fold
J_25-50 (- 89)	>25 & <50	-	10,703 (11,047)	5-fold
J_25-50_Laag (- 89)	>25 & <50	Low	10,871 (11,043)	5-fold
J_25-50_Midden(- 89)	>25 & <50	Medium	10,002 (11,074)	5-fold
J_25-50_Hoog (- 89)	>25 & <50	High	10,709 (11,270)	5-fold

Table 4.6: Different samples used to train models

As discussed in the data transformation section, redefining the unemployment period was necessary because some *catwz_uit* values were not useful in presenting 'job found'. However, one can argue about those correct proxies in the sense that a person who has found a job for one day is not actually employed. In line with this argument, all records with temporary work (value for *catwz_uit* equals 89) were removed from the dataset and new samples were drawn. The number of records in these samples is presented between brackets in Table 4.6. Class distribution of *All (-89)* sample is 33.46% High, 19.80% Medium, and 46.74% Low. Similar to the *All* sample, the sample size of *All (-89)* includes 10% of the population without 89.

4.3.3 Evaluation Metrics

Throughout this research, different evaluation measures were used. First of all, optimal parameters were chosen based on Cohen's kappa statistic. The 'caret' package in R has two embedded performance measures when dealing with multiclass labels: accuracy and kappa statistic. Accuracy could not be used due to imbalanced class distributions in the different samples. In contrast, Cohen's kappa statistic, which integrates the expected accuracy, is better in addressing the imbalanced problem [45]. As a result, it was preferred over accuracy. Note that a possible solution to cope with imbalanced class distributions is over or under sampling. However, due to a certain bias that evolves out of this solution, it was not preferred.

Next, precision[1], recall, and macro-averaged F score were computed for the final model of each machine learning technique. These performance measures are recommended for multiclass classification models, as discussed in M. Sokolova & G. Lapalme (2009) [74]. Macro-averaged F score was preferred over micro-averaged F score because the latter tends to be biased by imbalanced classes, as described by A. zgr et al (2005) [75] and Y. Yang & X. Liu (1999) [76]. As a result, the former was used to rank the models: the higher the metric, the better. Note that precision was often NaN, leading to an F score of NaN.

To round it off, all used metrics in the context of a multiclass classification task are defined below:

Cohen's Kappa Statistic measures the inter-rater agreement of different class labels. The raters can be seen as the columns and rows of a confusion matrix (actual vs predicted). Both observed accuracy and expected accuracy (random classifier) are computed. As a result, one can calculate Kappa statistic as the difference in observed and expected accuracy over one minus the expected accuracy. More information about Kappa statistic and its scaling interpretation can be found in A.J. Viera & J.M. Garrett (2005) [77].

Precision defines the number of correctly classified instances of a certain class label compared to the number of instances that were predicted for that class label [74].

Recall defines the number of correctly classified instances of a certain label compared to the number of instances for that label in the entire dataset [74].

Macro-averaged F score is a combination of precision and recall, as discussed in M. Sokolova & G. Lapalme (2009) [74]. Note that from a business context, preference is given to precision.

[1]Precision or positive predictive values = PPV

Chapter 5

Results

Results are presented in the following subsections. First, the variable importance according to the Cox proportional hazard model is presented. Second, the two general samples, with and without temporary work, are discussed based on their evaluation measures. Next, the results of subsamples as tabulated in Table 4.6 are discussed. Finally, the results of models without temporary work are illustrated.

5.1 General Samples

5.1.1 Cox PH Model

The Cox proportional hazard model was used to plot a survival graph of both samples in order to portray the relation between the number of unemployed days and the probability of finding a job. This graph is shown in Figure 5.1. Note that censored cases were not included as the curve would be rendered unreadable.

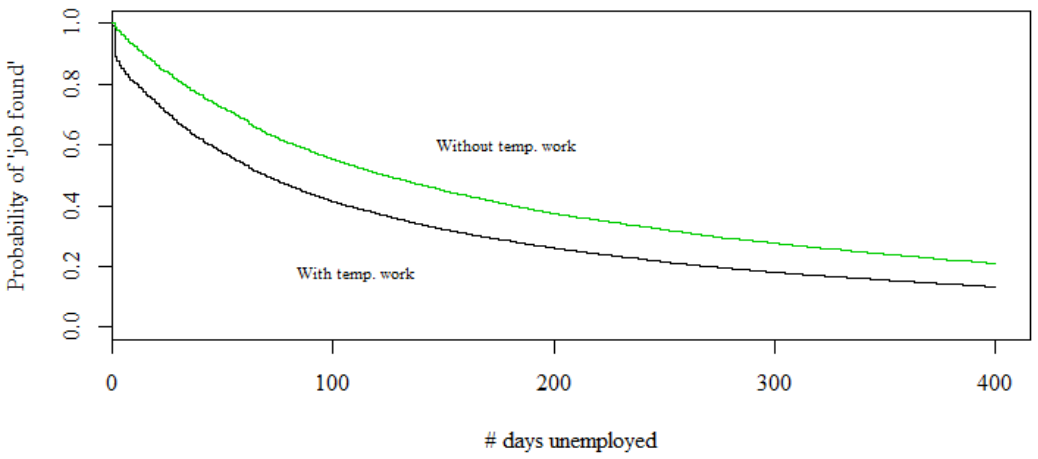

Figure 5.1: Survival curve of both samples

It makes sense that the curve with temporary work is lower than that without because a lot of records with temporary work found a job after 1-10 days. As a result, the curve lowered 'faster' than the one without temporary work. In addition, results from both curves are useless to compare with each other. Besides the survival graph, the most significant attributes in the cox model were also computed. In Table 5.1, the top eight (p-value <2e-16, α of 5%) are illustrated for both with and without temporary work. Note that most attributes are the same in both samples, which suggests there is not much of a difference in fitting both. More extensive results from the cox regression model for the general sample are attached in appendix 7, but will not be

discussed further. Hence, we refer to V. Ciuca & M. Matei (2010) [7] for a deeper analysis with Cox PH model.

All with temporary work	All without temporary work
Catwz_in_2/5/11 (base 0)	Catwz_in_2 (base 0)
Leeftijd_start_wklshd	Leeftijd_start_wklshd
Provincie_west_vl (base Antwerpen)	DGRAGH_N (base J)
DGRAGH_N (base J)	Studie_niveau_laag/midden (base hoog)
Taal_N_NANA (base 0)	Taal_N_0/1/2/3 (base NaNa)
Wrkls_periodes_voor_instr	Dagen_wrkls_10j_voor
Dagen_wrkls_10j_voor	Aantal_bedrijven
Aantal_bedrijven	DGRALL_N (base J)

Table 5.1: Significant attributes

5.1.2 Predictive Models

Turning to the results of different machine learning techniques in both samples, precision (PPV), recall, and the macro-averaged F score are tabulated in Tables 5.2 and 5.3. More precisely, precision and recall of each class label - High, Medium, Low - are shown together with the overall F score of the classifier. The best classifier is chosen based on the F score, as mentioned in 4.3.3 on evaluation metrics.

All	High (25%)		Medium (16%)		Low (58%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5634	0.3628	NaN	0.0000	0.6446	0.9269	NaN
Bagging RPART	0.5487	0.3210	NaN	0.0000	0.6369	0.9319	NaN
RandomForest	0.5922	0.3791	0.3415	0.0104	0.6493	0.9292	0.4156
J48	0.5092	0.3780	0.2626	0.0542	0.6523	0.8713	0.4236
GBM	0.5700	0.4543	0.2840	0.0171	0.6657	0.9021	0.4364
C5.0	0.5666	0.4307	0.2292	0.0082	0.6608	0.9106	0.4237
JRip	0.5674	0.3417	NaN	0.0000	0.6363	0.9264	NaN
Nnet	0.4994	0.4071	NaN	0.0000	0.6448	0.8797	NaN
SVM	0.5200	0.3648	0.2344	0.0223	0.6427	0.8909	0.4054
NB	0.1818	0.0010	0.5000	0.0030	0.5830	0.9975	0.2479
OLR	0.5290	0.3097	NaN	0.0000	0.6300	0.9218	NaN

Table 5.2: Results from general sample with temporary work

For every algorithm, recall and precision of the class label Low are the highest compared to the other two labels. Recall of Medium is at most 0.0542, which is extremely low. In addition, five out of eleven techniques give a precision of NaN. These NaN values in the PPV columns mean that there were no predictions for a certain label, e.g. if there were 1000 instances with class label Medium, then none of them was actually classified as Medium. Most of the techniques have a precision above 60% for the Low label and a recall for the Low label that is about twice that of the High label, expect for Naive Bayes. It is clear that Naive Bayes is unable to distinguish High class labels, resulting in an F score of 0.2479. In general, the best classifiers are C5.0, J48 and gradient boosting, with respective F scores of 0.4237, 0.4236 and 0.4364. Note that C5.0 and J48 are similar algorithms. Significance tests are performed in section 6.1.

Next, the results for the sample without temporary work are discussed in Table 5.3. In line with the previous sample, the combination of recall and precision for the Low class label is again for every technique the highest compared to the other two labels. However, they are lower than the previous sample. Medium has again very low support, not even covering 10%. Precision of High

is on average 0.5296, which is almost equal to that of the Low label (0.5675). Overall, the best classifiers are again J48, C5.0 and gradient boosting (GBM), with respective macro-averaged F scores of 0.4412, 0.4314 and 0.4355. However, these are still quite low scores compared to the highest possible value of 1.

All (-89)	High (33%)		Medium (19%)		Low (47%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5362	0.5894	0.5227	0.0045	0.5750	0.7757	0.4103
Bagging RPART	0.5221	0.5658	NaN	0,0000	0.5671	0.7734	NaN
RandomForest	0.5740	0.5808	0.3478	0.0109	0.5842	0.8185	0.4268
J48	0.5326	0.5778	0.2546	0.0811	0.5863	0.7201	0.4412
GBM	0.5772	0.6183	0.3228	0.0119	0.5963	0.8091	0.4355
C5.0	0.5526	0.6169	0.3875	0.0181	0.5943	0.7851	0.4314
JRip	0.6038	0.4281	0.5218	0.0039	0.5421	0.8806	0.3933
Nnet	0.5294	0.6227	NaN	0.0000	0.5896	0.7657	NaN
SVM	0.5546	0.5687	0.2037	0.0021	0.5762	0.8068	0.4129
NB	0.3000	0.0007	0,0000	0.0000	0.4699	0.9987	NaN
OLR	0.5431	0.5176	NaN	0.0000	0.5617	0.8178	NaN

Table 5.3: Results from general sample without temporary work

All things considered, Low class labels are easier to predict than Medium class labels and to a lesser extent High class labels. In addition, GBM is the best albeit still underperforming classifier over the two samples (with and without 89). These underwhelming results suggest that classifiers are not capable of fitting the data well. In the following sections, models' performances on the subsamples are discussed.

5.2 Subsamples with Temporary Work

In this section, results regarding certain subsamples including temporary work are discussed. We refer to Table 4.6 for the names and specifications of those samples. Samples' results were chosen based on the business context. More precisely, acquiring more information about lowly educated people under the age of 26 and highly educated people above the age of 49 was interesting. The other results can be found in appendix 4.

J_25_Laag	High (30%)		Medium (17%)		Low (52%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5246	0.4933	0.4151	0.0369	0.6227	0.8303	0.4229
Bagging RPART	0.5254	0.4480	NaN	0.0000	0.6084	0.8595	NaN
RandomForest	0.5688	0.4740	0.3438	0.0184	0.6177	0.8689	0.4383
J48	0.5025	0.3969	0.2312	0.0725	0.5978	0.8044	0.4548
GBM	0.5527	0.4798	0.2292	0.0185	0.6156	0.8479	0.4571
C5.0	0.5283	0.4942	0.3696	0.0285	0.6250	0.8375	0.4278
JRip	0.5836	0.3902	NaN	0.0000	0.5920	0.8986	NaN
Nnet	0.5215	0.4316	NaN	0.0000	0.6023	0.8595	NaN
SVM	0.5015	0.4827	0.1270	0.0134	0.6112	0.8039	0.4049
NB	NaN	0.0000	0.0000	0.0000	0.5257	0.9966	NaN
OLR	0.5011	0.4403	NaN	0.0000	0.5972	0.8347	NaN

Table 5.4: Results from sample J_25_Laag

Performances shown in Table 5.4 are not very different from the general sample. However, recall has lowered slightly for the Low label and increased somewhat for the High label. Medium class is still very hard to fit and GBM is again among the best classifiers together with J48 and to a lesser extent Random forest.

J_50_Hoog	High (38%)		Medium (16%)		Low (46%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5685	0.6278	0.2222	0.0038	0.6080	0.7723	0.4281
Bagging RPART	0.5396	0.6385	NaN	0.0000	0.6096	0.7380	NaN
RandomForest	0.5697	0.6450	0.0909	0.0132	0.6160	0.7394	0.4334
J48	0.4877	0.6661	0.1954	0.0644	0.6157	0.5772	0.4186
GBM	0.5756	0.6352	0.1875	0.0113	0.6137	0.7702	0.4361
C5.0	0.5683	0.6122	0.0000	0.0000	0.6014	0.7791	NaN
JRip	0.5960	0.5127	NaN	0.0000	0.5644	0.8354	NaN
Nnet	0.5365	0.6763	NaN	0.0000	0.6176	0.7078	NaN
SVM	0.5304	0.6516	0.2667	0.0075	0.6159	0.7160	0.4206
NB	0.4601	0.7666	0.0000	0.0000	0.6395	0.5158	NaN
OLR	0.5268	0.6621	NaN	0.0000	0.6123	0.7652	NaN

Table 5.5: Results from sample J_50_Hoog

Similar findings hold for results in Table 5.5. Random forest and GBM perform just about the same, with an almost identical F score as consequence. Again, all learners have difficulties with fitting the medium class label. Notice that the recall of the Low label has lowered with nearly 10% on average. In other words, when comparing highly educated people above the age of 49 with low unemployment duration to lowly educated people under the age of 26 with low unemployment duration, the former was more difficult to support as such.

To summarize, the overall F scores of different data mining techniques still underperform. A lot of them are incapable of predicting the medium class labels, where the highest recall on both samples is 10.81%. In contrast, the Low and to a lesser extent High interval are better predicted, meaning that people at risk of long-term unemployment are distinguished from those with shorter term (0-3 months). For example, according to GBM (Table 5.5), 77% of the Low instances were recognized as short-term unemployed, and 61% of the records predicted as short-term unemployed were correctly classified as such.

5.3 Subsamples without Temporary Work

In contrast to the two samples in the previous section, temporary work is now excluded. Table 5.6 presents results for lowly educated people under the age of 26, while Table 5.7 shows those for highly educated people above the age of 49. More results are shown in appendix 5.

J_25_Laag(-89)	High (40%)		Medium (21%)		Low (39%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5905	0.6574	0.2471	0.0324	0.5159	0.6872	0.4229
Bagging RPART	0.5764	0.7015	NaN	0.0000	0.5265	0.6775	NaN
RandomForest	0.6089	0.7051	0.3512	0.0262	0.5396	0.7084	0.4383
J48	0.5867	0.6645	0.2979	0.1081	0.5363	0.6374	0.4548
GBM	0.6176	0.6912	0.3737	0.0571	0.5457	0.7173	0.4571
C5.0	0.6141	0.7022	0.2258	0.0108	0.5299	0.7116	0.4278
JRip	0.5356	0.8125	NaN	0.0000	0.5663	0.5410	NaN
Nnet	0.5552	0.7279	NaN	0.0000	0.5220	0.6174	NaN
SVM	0.5582	0.7125	0.2333	0.0108	0.5214	0.6239	0.4049
NB	0.6892	0.3228	0.0000	0.0000	0.4356	0.9204	NaN
OLR	0.5700	0.7007	NaN	0.0000	0.5137	0.6539	NaN

Table 5.6: Results from sample J_25_Laag (-89)

Records with long unemployment duration are predicted in quite the same manner as those with short unemployment duration. The average precision and recall are respectively 0.5404 and 0.6726 for the High label. The Medium class is still very difficult to fit, with barely 10% as highest recall and about 40% as highest precision. J48 and GBM are the best classifiers.

J_50_Hoog(-89)	High (46%)		Medium (19%)		Low (35%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5727	0.7885	0.3500	0.0113	0.5197	0.5311	0.4036
Bagging RPART	0.5661	0.7951	NaN	0.0000	0.5140	0.5153	NaN
RandomForest	0.5885	0.8281	0.2125	0.0276	0.5604	0.5206	0.4255
J48	0.5237	0.7300	0.1644	0.0797	0.4681	0.3532	0.3733
GBM	0.5960	0.8221	0.2826	0.0211	0.5563	0.5539	0.4285
C5.0	0.5703	0.8096	0.2667	0.0130	0.5445	0.5206	NaN
JRip	0.5215	0.9124	NaN	0.0000	0.5871	0.3190	NaN
Nnet	0.5641	0.7938	NaN	0.0000	0.5202	0.5197	NaN
SVM	0.5661	0.8182	0.1628	0.0113	0.5399	0.4917	0.4017
NB	0.4648	0.9947	0.0000	0.0000	0.5600	0.0123	NaN
OLR	0.5394	0.8038	NaN	0.0000	0.5180	0.4654	NaN

Table 5.7: Results from sample J_50_Hoog (-89)

The High interval is now predicted quite well in terms of recall, as shown in Table 5.7. However, Medium intervals are still poorly fitted. The Low class label's recall lowered with 15% to 20% on average. Precision on the High and Low label are nearly the same, with an average of around 0.54. On the whole, gradient boosting and Random forest seem to be the best in predicting unemployment duration. Note that both models highly underperform with respective F scores of 0.4285 and 0.4255.

To summarize, distinguishing records from Medium class labels still poses a challenge. This suggests that some further research could be done with regard to the different interval thresholds. Moreover, one can train models on even smaller subgroups, e.g. people under the age of 26 with a degree and driver's license. In this way, classifiers might fit the data better and good predictive models could be built. Unfortunately, we were unable to conduct further research on this issue due to time constraints.

Chapter 6

Benchmarking

Relationships between data mining techniques and samples are discussed. First, the best and worst performing algorithms are illustrated and compared over both kinds of samples. Next, the macro-averaged F scores of the top three algorithms are depicted on the basis of the different samples. A graph is also plotted to discuss the differences in performance between the white and black box techniques. Finally, a sneak peek is provided for further research.

6.1 Comparing Data Mining Techniques

In Figure 6.1, the top scoring data mining techniques are illustrated. The absolute values in the pie charts represent the number of times a certain technique occurred in a sample's top three. The top three represent the three highest F scores over all samples with and without temporary work. More precisely, the pie chart to the left in Figure 6.1 presents the best classifiers over all samples, while that to the right presents those over all samples without temporary work.

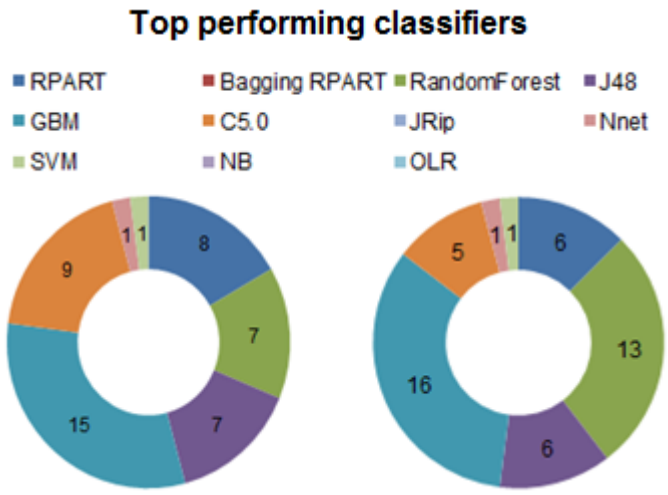

Figure 6.1: Classifiers over all samples (left), without temporary work (right)

There is one main difference between the two pie charts. Random forest (13) is well represented among the best classifiers on the right, while on the left it makes way for C5.0 decision tree (9). Similar in both charts is the fact that artificial neural networks and support vector machines each appear only once. It seems that Random forest is better at fitting the data when the lower boundary of number of days until employment is higher. On the other hand, when more instances have a very low number of days until employment, C5.0 and to a lesser extent Rpart are the better learners. Note that some people find a temporary job after two days, and a 'real' (full- or part-time) job only after 100 days. Generally speaking, gradient boosting is the best classifier.

The worst learners are depicted in Figure 6.2. The absolute values in the pie charts represent the number of times a certain technique occurred with the lowest F score of a sample.

Every time, more than three algorithms had NaN values. As a result, they all have the same (lowest) value. Notice that on the left chart the total number of NaN occurrences (84) is higher than that on the right chart (72). In other words, more classifiers had difficulties in fitting the samples with temporary work.

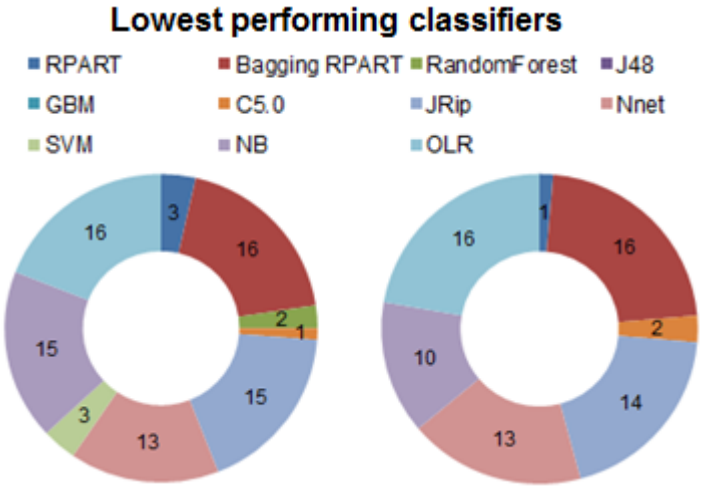

Figure 6.2: Classifiers over all samples (left), without temporary work (right)

On the whole, Rpart bagging, Ordinal logistic regression, and JRip are the worst performers. From the pie chart on the left, Naive Bayes and Nnet also have many NaN values. While on the right chart, support vector machines is not represented in the top three algorithms nor is it found among the poorest learners. Notice that on the left chart, J48 and GBM have zero occurrences, while on the right chart Random forest, GBM, J48 and support vector machines have zero occurrences.

Next, the top three performing data mining techniques are depicted in detail in Figure 6.3. More precisely, each sample with its respective F score is illustrated for the three methods. Notice that for each sample the F scores of the three top performing techniques are shown, even if they were not among the best three for a specific sample. NaN values are excluded and represented by a discontinuous interval. Finally, J48 was preferred over Rpart because the latter had two non-values.

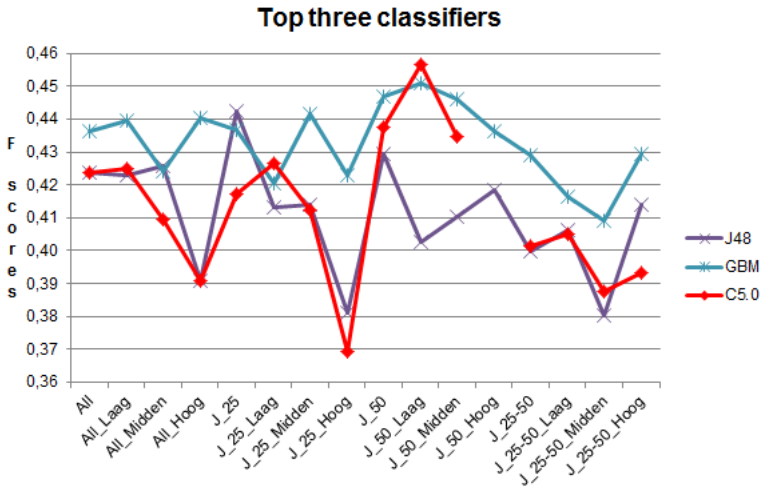

Figure 6.3: Top three performing classifiers

In general, GBM seems to be better in predicting new instances than C5.0 and J48 decision tree. The highest F scores can be found in samples with age above 50. In contrast, the lowest overall performances are found in the sample with highly educated people under the age of 26: *All_25_Hoog*. In Figure 6.4, the same is depicted for samples without temporary work. Notice that Random forest, GBM, and J48 were the best classifiers.

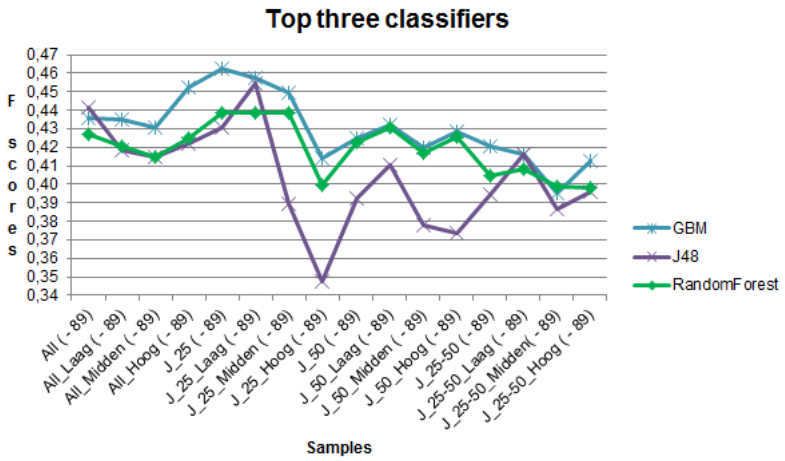

Figure 6.4: Top three performing classifiers (-89)

On the whole, Random forest and GBM are better in fitting the data and thus predicting new instances. The performances of J48 decision tree are high at the *J_25_Laag* sample but subsequently dip to very low at *J_25_Hoog*. GBM is able to maintain a nearly constant performance with heights in samples where age is under 26. The same is true of Random forest. Comparing both figures, it would seem that samples with age above 50 are better classified when temporary work is included, while instances under the age of 26 are better predicted when temporary work is excluded. However, if the three classifiers in each figure are combined into their respective averages per sample, then there is no significant difference between those averages, as shown by a paired Wilcoxon signed-rank test (p-value of 0.978, 5% α). To conclude, we again want to stress the fact that F scores around 0.4 are actually well below par.

Finally, white box versus black box techniques are discussed. As already mentioned in the introduction, it is important for the VDAB to have some explanatory indications of the final output. White box algorithms are in a sense more interpretable than black box methods because their structure gives insights into the final result. Two of the eleven techniques are considered as black box algorithms: neural networks and support vector machines. The others are white box methods, even gradient boosting machines according to Natekin and Knoll (2013) [78].

In Figure 6.5, support vector machines are compared to gradient boosting machines. Gradient boosting is the best classifier over all samples, and SVM is a black box technique whose performance fluctuates between high and low. It makes no sense to compare them with neural networks as they have NaN values on nearly all samples. We can see that support vector machines perform reasonably well compared to gradient boosting on all samples except three: *All_Laag*, *J_50_Laag*, and *J_50_Midden*. Nevertheless, a paired Wilcoxon signed-rank test ($\alpha = 5\%$) shows that on average GBM performs significantly better than SVM: p-value of 2.4410e-04. This kind of test was preferable to a normal t-test because the F scores were not normally distributed.

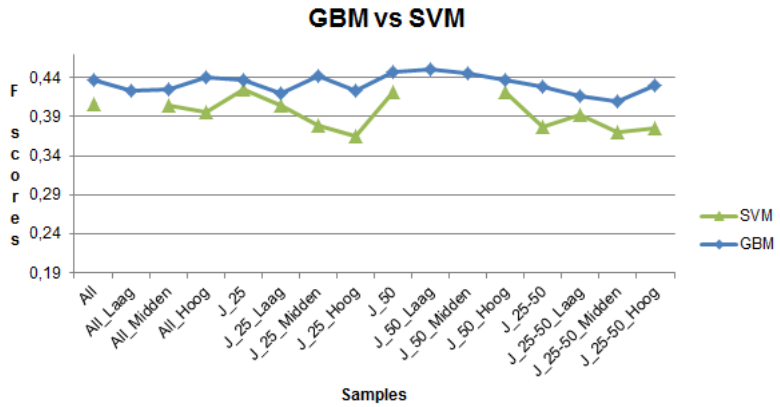

Figure 6.5: Random forest compared to support vector machines on all samples

In Figure 6.6, Random forest, SVM and GBM are depicted. GBM is on average the best white box classifier on samples without temporary work. It performs on average significantly better than support vector machines (p-value of 3.052e-05, 5% α) and to a lesser extent than random forest (p-value of 3.052e-04, 5% α). In addition, random forest performs also significantly better than SVM (p-value of 3.052e-04, 5% α).

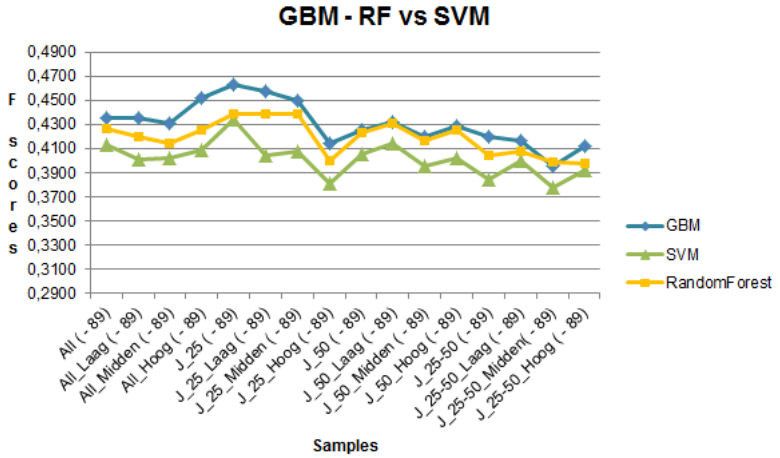

Figure 6.6: GBM and RF compared to SVM

In conclusion, white box algorithms - especially GBM - are on average significantly better in predicting new instances compared to black box methods on both kinds of samples.

6.2 Comparing Time Intervals

Some further research into the different labelling could be conducted. Here, we try to give a brief introduction on how different time intervals can change the performance of six machine learning techniques. More precisely, a combination of the best and worst methods was chosen. The *All* sample and *All sample (-89)* are used to train and test the algorithms, and are shown respectively in Figures 6.7 and 6.8. More detailed results are tabulated in appendix 6, where 'High' is set counterintuitively as the 'positive' class label. Samples are drawn randomly (~10,000 records)

and split into 70/30% for training and testing models. Class distributions are also tabulated in appendix 6 where H stands for High and L for Low.

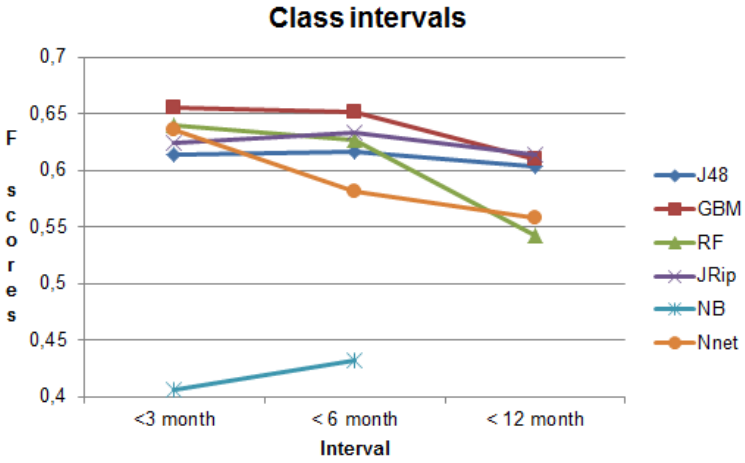

Figure 6.7: Comparison of different time intervals

Three different time intervals are used; more specifically, the split between the Low and High label is made consecutively after 3, 6, and 12 months. Notice that these are binary class labels instead of multiclass ones, which were initially used in this research. The medium label was always fitted very poorly, so it was removed in order to have better distinguishable groups. We can see that overall the F scores are improved to an average of around 0.6, excluding Naive Bayes. All algorithms considered, there are no significant differences among the time intervals (p-value between <3 & <6 months equals 0.8438, <6 & <12 p-value of 0.0625 and <3 & <12 p-value of 0.0625 5% α).

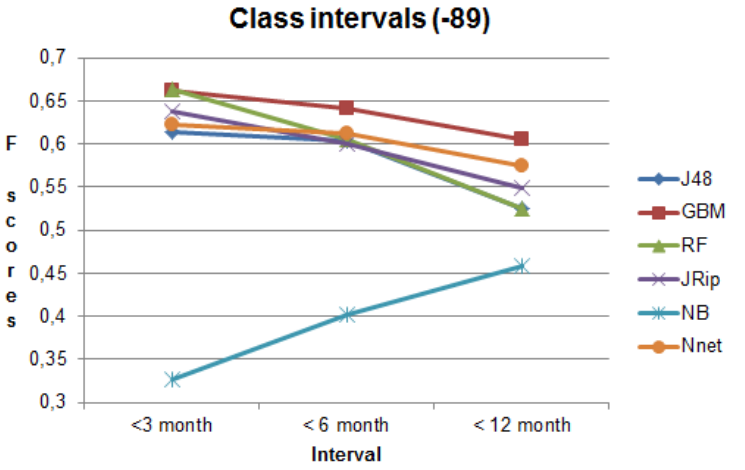

Figure 6.8: Comparison of different time intervals in sample without temporary work

F scores are also around 0.6, excluding Naive Bayes. Performances over intervals are not significantly different, all algorithms considered (p-value between <3 & <6 months equals 0.4375, <6 & <12 p-value of 0.2188 and <3 & <12 p-value of 0.3125 5% α). Naive Bayes is for both samples the most underperforming classifier.

Chapter 7

Recommendations to the VDAB

In this thesis, different prediction models were trained to predict at registration the unemployment duration of a job seeker on the basis of his/her distinctive personal characteristics. In addition, most of these models were found capable of providing insights as to why certain job seekers are at risk, e.g. lowly educated. These outcomes can be used by the VDAB to become more proactive in their effort to provide guidance to job seekers.

The 'best' models for each sample are indicated in Table 7.1. These models were chosen based on their performance with regard to precision and recall of the Low and High label. The Medium label was very hard to predict and is less important when the other two are well distinguishable. Only samples that include temporary work are illustrated because no significant difference was found between samples with and without temporary work (see section 6.1). Note that the macro-averaged F score is still computed over all three class labels.

Sample	Technique	High		Medium		Low		Overall
		PPV	Recall	PPV	Recall	PPV	Recall	MacroF
All	GBM	0.5700	0.4543	0.2840	0.0171	0.6657	0.9021	0.4364
All_Laag	RF	0.6025	0.4856	0.2500	0.0079	0.6404	0.8893	0.4325
All_Midden	GBM	0.5422	0.4070	0.1837	0.0166	0.6795	0.9076	0.4242
All_Hoog	Nnet	0.4499	0.4831	NaN	0.0000	0.6691	0.8653	NaN
J_25	GBM	0.5979	0.4018	0.2432	0.0163	0.6922	0.9466	0.4369
J_25_Laag	RF	0.5688	0.4740	0.3438	0.0184	0.6177	0.8689	0.4247
J_25_Midden	GBM	0.5229	0.3404	0.3426	0.0634	0.7129	0.9260	0.4416
J_25_Hoog	GBM	0.4658	0.2086	0.3362	0.1119	0.7260	0.9239	0.4231
J_50	GBM	0.5852	0.6134	0.5000	0.0089	0.6488	0.8189	0.4468
J_50_Laag	GBM	0.5938	0.6032	0.1094	0.0156	0.6598	0.8090	0.4509
J_50_Midden	GBM	0.5998	0.5752	0.3077	0.0087	0.6487	0.8457	0.4461
J_50_Hoog	GBM	0.5756	0.6352	0.1875	0.0113	0.6137	0.7702	0.4361
J_25-50	GBM	0.5541	0.4249	0.2881	0.0331	0.6457	0.8829	0.4288
J_25-50_Laag	GBM	0.5592	0.4615	0.1111	0.0021	0.6432	0.8715	0.4166
J_25-50_Midden	GBM	0.4945	0.3990	0.2553	0.0253	0.6476	0.8623	0.4091
J_25-50_Hoog	GBM	0.5325	0.4467	0.3000	0.0268	0.6556	0.8861	0.4296

Table 7.1: Results based on Low & High class label

Based on the sample *All_Laag*, the random forest model has the highest precision of the High label over all samples, as illustrated in Table 7.1. Put alternatively: if a lowly educated person is predicted as long-term unemployed, then in 60.25% of cases a prediction as such will be accurate. On the other hand, the gradient boosting model, based on the sample *J_25_Hoog*, has the highest precision of the Low label. This means that if a highly educated person under the age of 26 is predicted as short-term unemployed, then in 72.60% of cases such a prediction will be spot on.

Finally, some recommendations regarding further research may here be proffered. One of the limitations of the current research was a lack of attributes describing job seekers' motivation, willingness to work, social and soft skills (e.g. some people are extrovert, while others are introvert). These attributes might have a significant impact on the risk of long-term unemployment

and could distinguish between different groups in terms of unemployment duration [79]. Further research could look into ways of incorporating these attributes as well. This, however, might come with possibly inaccurate reflections of reality due to the less quantifiable and more subjective nature of this kind of attributes.

In addition, further research could also examine how models behave on more detailed samples. In the current research, samples were filtered based on two attributes: age and education. However, another attribute diversifier such as driving license could also be introduced. Or, more detailed information about the degree of education/previous job experience might be proposed, e.g. train models only on a sample of engineers or fishermen living relatively far away from a harbour. In this way, a variety of small models might better fit the data and hence have a better predictive performance. Note that in this research very detailed attributes were aggregated, e.g. level of education or city/town. As education and residence are significant attributes in explaining the unemployment duration, one could use them in a more detailed manner. However, this kind of fine granularity - 1196 factor levels in Studies - will take a certain amount of time to model.

Chapter 8

Conclusion

The main objective was to predict and explain the unemployment duration among the newly unemployed so as to provide better guidance to those who are at risk of long-term unemployment. Different machine learning techniques were tested to construct predictive models upon multiple specified samples. These samples were split into 70/30% to train and subsequently test the performance of the model. Generally speaking, the performance of those models in terms of F scores was rather low, with the highest score at around 0.46, which is quite low compared to the highest possible value of 1.

In all, gradient boosting performs best in predicting unemployment duration. However, for certain groups of job seekers, unemployment duration is best predicted using the J48 algorithm, C5.0 decision tree or Random Forest. On the whole, the performance of classifiers on samples without temporary work is lower than that of those with such work. This is mostly due to a better recall of the Medium class label in the latter scenario. On the one hand, classifiers tend to be better at predicting unemployment duration for job seekers under the age of 26 when temporary work is not included in the train data. While on the other hand, job seekers above the age of 49 are better distinguished from the other samples when temporary work is included in the train data. However, all samples considered, the difference in performance found between samples with and without temporary work is not significant.

As described in the introduction, the high number of people in long-term unemployment is a key issue. It is crucial to provide better guidance to job seekers and work more proactively towards significantly reducing the number of long-term unemployment. This research can thus conclude that short-term unemployment and to a lesser extent long-term unemployment are both easier to predict than medium-term unemployment. This means that people at risk of long-term unemployment (more than 6 months) are distinguishable from those with short-term unemployment (up to 3 months). As a result, it is possible to provide guidance in a more efficient manner to the former group. However, any conclusion based on these results should be handled with caution as the predictive performance is low.

Another important finding of this research is that in general white box algorithms performed significantly better in predicting unemployment duration than black box methods. This means that it is possible to offer insights into the predicted unemployment duration by means of explanatory indications. This allows for addressing job seekers' characteristics that are negatively impacting their unemployment duration, thereby improving their chance of finding a job in a more timely manner.

A brief comparison of different time intervals has shown that classifiers better fit the data when the medium class label is removed. Excluding Naive Bayes, F scores rose from an average of 0.4 to 0.6. In addition, there are no significant differences between different splits in interval. However, when Naive Bayes is excluded, prediction models perform better when the Low class is split from the high class label at three months. Nevertheless, considering three months as a fair threshold for long-term unemployment is not a straightforward decision.

Finally, further research paths include building models on more detailed samples by raising the number of filters, adding more detailed information on certain aggregated attributes, and incorporating attributes that describe job seekers' social and soft skills, motivation, and willingness to work.

Appendix A

1. List of attributes

Attribute	Exclude	Reason to exclude	Explanation
KLNR	yes	No explanatory use	ID number of a person
DATUM_INSTROOM_WKLSHD	yes	No explanatory use	Date of signing up at VDAB
DATUM_UITSTROOM_WKLSHD	yes	No explanatory use	Date of signing out at VDAB
AANTAL_DAGEN_WERKLOOS	yes	Changed to tijdsinterval (Class.)	Difference between Date_in and Date_out
CATWZ_IN(STROOM)	no		Code of signing up at the VDAB
CATWZ_UIT(STROOM)	yes	No explanatory use	Code of signing out at the VDAB
LEEFTIJD_START_WKLSHD	no		Age at start of unemployment (14<x<67)
PROVINCIE	no		Province of accommodation
GEMEENTE	yes	Aggregated into Provincie	Town of accommodation
CD_POST	yes	Equivalent to Gemeente	City code
CD_NIS	yes	Equivalent to CD_POST	NIS code
DGRLFT	yes	Included in leeftijd_start_wklshd	Age at start of unemployment
DGRAGH	no		Disabled
DGRLGS	yes	Aggregated into studie_niveau	Lowly educated
DGRVER	yes	Constant value	/
DGRALL	no		Foreign
DGRZAG	yes	Constant value	Heavily disabled
EIGEN_WAGEN	no		Owns car
RIJBEWIJS	no		Drivers license
CAT_ARBGESCH	no		Able to work
GESLACHT	no		Gender
NATIONALITEIT	no	Aggregated into Continent	Nationality
GEBOORTEDAT	yes	Included in leeftijd_start_wklshd	Date of birth
STUDNIV_DBDA	yes	Aggregated into studie_niveau	Education ranking
UITREIKINGSDATUM_LED	yes	Too detailed	Date reception degree
AFSTUDEER_JAAR_BURGER	yes	Missing values (+-50%)	Graduating date
STUDIE_NIVEAU	no		Education level (low, medium, high)
STUDIES	yes	Aggregated into studie_niveau	Degree of studies
STUDIES2	yes	Aggregated into studie_niveau	Extra degree of studies

STUDIESA	yes	Aggregated into studie_niveau	Extra degree of studies
KENNIS_NEDERLANDS	yes	Included in Taal_N	Knowledge of Dutch
TAAL_N	no		Knowledge of Dutch
TAAL_F	no		Knowledge of French
TAAL_E	no		Knowledge of English
MOEDERTAAL	yes	No explanatory use	Mother tongue
AANTAL_TALEN_NIV_1	yes	Included in Taal_N,F,E	Languages of level 1
AANTAL_TALEN_NIV_2	yes	Included in Taal_N,F,E	Languages of level 2
AANTAL_TALEN_NIV_3	yes	Included in Taal_N,F,E	Languages of level 3
WRKLS_PERIODES_VOOR_INSTR	no		Number of periods unemployed
DAGEN_WERK_10J_VOOR	yes	Aggregated into Arbeid_dagen	# worked days within 10 years
DAGEN_WRKLS_10J_VOOR	no		# days unemployed within 10 years
DAGEN_WERK_1J_NA	yes	Constant value	# days worked one year after signing up
DAGEN_WRKLS_1J_NA	yes	No explanatory use	# days unemployed one year after signing up
AANTAL_MOM	yes	No explanatory use	# tailored emails
AANTAL_MOM_1J_VOOR	yes	No explanatory use	# tailored emails one year before job
AM	yes	No explanatory use	# matching job openings
AM_MAIL_TOEGEKOMEN	yes	No explanatory use	# matching job openings received by mail
AM_GELEZEN	yes	No explanatory use	# matching job openings read
AM_1J_VOOR	yes	No explanatory use	# matching job openings one year before job
AM_MAIL_TOEGEKOMEN_1J_VOOR	yes	No explanatory use	# matching job openings received by mail one year before job
AM_GELEZEN_1J_VOOR	yes	No explanatory use	# matching job openings read one year before job
CVS_INSTROOM	no		# CVs
CVS_TOEGEVOEGD	yes	No explanatory use	# CVs added
GEWENSTE_JOBS_INSTROOM	no		# wanted jobs
GEW_JOBS_TOEGEVOEGD	yes	No explanatory use	# wanted jobs added
GEWENSTE_JOBS_ERV_1_INSTROOM	no		# wanted jobs with one-year experience
GEW_JOBS_ERV_1_TOEGEVOEGD	yes	No explanatory use	# wanted jobs with one-year experience added
GEWENSTE_JOBS_ERV_2_INSTROOM	no		# wanted jobs with two-year experience
GEW_JOBS_ERV_2_TOEGEVOEGD	yes	No explanatory use	# wanted jobs with two-year experience added
GEWENSTE_JOBS_ERV_3_INSTROOM	no		# wanted jobs with three-year experience
GEW_JOBS_ERV_3_TOEGEVOEGD	yes	No explanatory use	# wanted jobs with three-year experience added
GEWENSTE_JOBS_ERV_4_INSTROOM	no		# wanted jobs with four-year experience
GEW_JOBS_ERV_4_TOEGEVOEGD	yes	No explanatory use	# wanted jobs with four-year experience added
GEWENSTE_KNELPUNTBEROEPEN	no		# wanted bottleneck jobs

GEWENSTE_UITZONDERING_BEROEPEN	no		# wanted jobs with no openings
MAX_RATIO_VAC_PER_KANDIDATEN	no		Ratio job openings & # applicants
TOTAAL_AANTAL_VACATURES	yes	Included in max_ratio_vac..	# job openings
TOTAAL_AANTAL_KANDIDATEN	yes	Included in max_ratio_vac..	# job applicants
OPL_UREN_VOOR_INSTROOM	no		# training hours before unemployment
OPL_VOOR_INSTROOM	yes	Included in opl_uren_voor_instroom	# training before unemployment
OPL_BEEINDIGD_VOOR_INSTROOM	yes	Included in opl_uren_voor_instroom	# training ended before unemployment
OPL_WEBLEREN_VOOR_INSTROOM	yes	Included in opl_uren_voor_instroom	# online training before unemployment
OPL_UREN	yes	No explanatory use	Training hours during unemployment
OPL_BEGONNEN	yes	No explanatory use	# training started during unemployment
OPL_BEEINDIGD	yes	No explanatory use	# training ended during unemployment
OPL_WEBLEREN	yes	No explanatory use	# online training during unemployment
INTERESSES_INSTROOM	no	Changed into true/false	# interests
INTERESSES_TOEGEVOEGD	yes	No explanatory use	# interests added
INTERESSES_NU	yes	No explanatory use	# interests now
REFERS_INSTROOM	no	Changed into true/false	# references
REFERS_TOEGEVOEGD	yes	No explanatory use	# references added
AANTAL_BEDRIJVEN	no		# companies worked for
ARBEID_DAGEN	no		# days worked
AANTAL_ARBEIDSCONTRACTEN	no		# employment contracts
AANTAL_BEDRIJVEN_INTERIM	yes	High correlation (0,94) with aantal_bedrijven	# companies worked for as temporary emp.
ARBEID_DAGEN_INTERIM	no		# days worked as temporary emp.
AANTAL_ARBEIDSCONTR_INTERIM	yes	High correlation (0,96) with aantal_arbeidscontracten	# employment contracts as temporary emp.
AANTAL_BEDRIJVEN_STUDENT	yes	Aggregated into aantal_bedrijven	# companies worked for as student
ARBEID_DAGEN_STUDENT	yes	Aggregated into arbeid_dagen	# days worked as student
AANTAL_ARBEIDSCONTR_STUDENT	yes	Aggregated into aantal_arbeidscontracten	# employment contracts as student
AANTAL_BEDRIJVEN_X	yes	Aggregated into aantal_bedrijven	# agricultural & catering companies worked for
ARBEID_DAGEN_X	yes	Aggregated into arbeid_dagen	# days worked at companies X
AANTAL_ARBEIDSCONTR_X	yes	Aggregated into aantal_arbeidscontracten	# employment contracts at companies X
AANTAL_BEDRIJVEN_IBO	yes	Aggregated into aantal_bedrijven	# companies gave professional training
ARBEID_DAGEN_IBO	yes	Aggregated into arbeid_dagen	# days worked at companies IBO

AANTAL_ARBEIDSCONTR_IBO	yes	Aggregated into aantal_arbeidscontracten	# employment contracts at companies IBO
AANTAL_BEDRIJVEN_BOUW	yes	Aggregated into aantal_bedrijven	# construction companies worked for
ARBEID_DAGEN_BOUW	yes	Aggregated into arbeid_dagen	# days worked at companies BOUW
AANTAL_ARBEIDSCONTR_BOUW	yes	Aggregated into aantal_arbeidscontracten	# employment contracts at companies BOUW
AANTAL_BEDRIJVEN_OTHER	yes	Aggregated into aantal_bedrijven	# other companies worked for
ARBEID_DAGEN_OTHER	yes	Aggregated into arbeid_dagen	# days worked at companies OTHER
AANTAL_ARBEIDSCONTR_OTHER	yes	Aggregated into aantal_arbeidscontracten	# employment contracts at companies OTHER
TIJDSINTERVAL	no		class label (time interval)

Table A.1: List of all attributes

2. Catwz_in/uit

Catwz_in	
Code	Explanation
0	Fully unemployed, benefit eligibility
2	Job seeker (Article 36) in professional integration time
3	Free-registered job seeker, not working
5	Compulsorily registered O.C.M.W.
6	Registration due to the supervision of a person with a disability (maximum degree OV2)
11	Job seeker in part-time education: job seeker studying part-time or following linked training BUSO-OV3
14	Job seeker excluded from the right to benefits

Table A.2: Catwz_in codes

Catwz_uit		
Code	Found a job?	Explanation
0	No	Fully unemployed, benefit eligibility
2	No	Job seeker (Article 36) in professional integration time
3	No	Free-registered job seeker, not working
5	No	Compulsorily registered O.C.M.W.
6	No	Registration due to the supervision of a person with a disability (maximum degree OV2)
11	No	Job seeker in part-time education: job seeker studying part-time or following linked training BUSO-OV3
14	No	Job seeker excluded from the right to benefits
18	No	Campus enrolment, student who will finish his/her studies at the end of the academic year
19	No	Job student: wishes to work as a job student
25	Yes	Third employment circuit - full time: working and looking for another job
30	Yes	UWV exempt from registration due to PWA activities: working and looking for another job
32	Yes	Dependent RIZIV in preparation for employment:
33	Yes	(Candidate) work care assistant
66	No	Deceased
70	Yes	Regular full-time placement
76	No	Unenrolled due to sickness
77	No	Unenrolled due to resumption of studies
78	Yes	Unenrolled due to job

79	No	Regular unenrolment
80	Yes	Part-time employee with benefits: part-time employee receiving benefits and looking for another job
82	Yes	Job seeker (Article 36) in professional integration time, working part-time
85	No	Job seeker in individual vocational training
89	Yes/No	Regularly works as interim
90	Yes	Working full-time, voluntarily registered job seeker
91	Yes	Working, part-time student, job seekers: job seeker working part-time and studying part-time
92	No	Temporarily unemployed
93	Yes	Working part-time, voluntarily registered job seeker
96	No	Enrolment exemption due to family, social reasons
97	No	Unemployed exempt from registering as a job seeker because of studies or vocational training

Table A.3: Catwz_uit codes

3. Inappropriate catwz_uit

	Case 1	Case 2	Case 3
Inappropriate proxies	0,2,3,5,6,11,14,18,19, 66,76,77,79,85,92,96,97	0,2,3,5,6,11,14, 18,19,76,79,92,96	76,77,85,96,97

Table A.4: Inappropriate codes

4. Results on subsamples with temporary work

All_Laag	High (30%)		Medium (16%)		Low (54%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5561	0.4587	0.2500	0.0079	0.6285	0.8661	0.4155
Bagging RPART	0.5799	0.3612	NaN	0.0000	0.6065	0.9107	NaN
RandomForest	0.6025	0.4856	0.2500	0.0079	0.6404	0.8893	0.4325
J48	0.5078	0.4886	0.2045	0.0355	0.6377	0.8028	0.4231
GBM	0.5740	0.5095	0.2400	0.0236	0.6460	0.8548	0.4395
C5.0	0.5701	0.5015	0.3333	0.0020	0.6408	0.8678	0.4250
JRip	0.5689	0.3861	NaN	0.0000	0.6099	0.8966	NaN
Nnet	0.5818	0.4000	NaN	0.0000	0.6151	0.9011	NaN
SVM	0.5638	0.4179	0.0000	0.0000	0.6174	0.8836	NaN
NB	0.4286	0.0060	NaN	0.0000	0.5395	0.9966	NaN
OLR	0.5798	0.4159	NaN	0.0000	0.6145	0.8898	NaN

Table A.5: Results on All_Laag

All_Midden	High (23%)		Medium (16%)		Low (61%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5060	0.3811	0.2615	0.0313	0.6751	0.8959	0.4202
Bagging RPART	0.6314	0.2545	NaN	0.0000	0.6503	0.9697	NaN
RandomForest	0.5750	0.3566	0.3750	0.0055	0.6699	0.9413	0.4113
J48	0.4835	0.3786	0.2388	0.0589	0.6752	0.8656	0.4259
GBM	0.5422	0.4070	0.1837	0.0166	0.6795	0.9076	0.4242
C5.0	0.5573	0.3708	0.2000	0.0018	0.6701	0.9311	0.4094
JRip	0.5369	0.3385	NaN	0.0000	0.6626	0.9311	NaN
Nnet	0.5091	0.3243	NaN	0.0000	0.6557	0.9198	NaN
SVM	0.4513	0.3592	0.2000	0.0350	0.6663	0.8636	0.4039
NB	0.1000	0.0013	0.0000	0.0000	0.6085	0.9951	NaN
OLR	0.5544	0.2830	NaN	0.0000	0.6509	0.9443	NaN

Table A.6: Results on All_Midden

All_Hoog	High (21%)		Medium (19%)		Low (60%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.4800	0.3181	0.1966	0.0342	0.6614	0.9058	0.4018
Bagging RPART	0.5382	0.2592	NaN	0.0000	0.6469	0.9590	NaN
RandomForest	0.5802	0.3602	0.1786	0.0078	0.6569	0.9366	0.4105
J48	0.4353	0.2727	0.2000	0.0550	0.6539	0.8814	0.3910
GBM	0.5512	0.3490	0.2796	0.0773	0.6727	0.9058	0.4402
C5.0	0.5306	0.3063	0.1176	0.0030	0.6583	0.9506	0.3907
JRip	0.5849	0.1826	NaN	0.0000	0.6354	0.9765	NaN
Nnet	0.4499	0.4831	NaN	0.0000	0.6691	0.8653	NaN
SVM	0.4746	0.2754	0.2581	0.0475	0.6498	0.9049	0.3951
NB	0.3333	0.0059	0.0000	0.0000	0.6120	0.9967	NaN
OLR	0.4969	0.2327	NaN	0.0000	0.6406	0.9524	NaN

Table A.7: Results on All_Hoog

J_25	High (19%)		Medium (18%)		Low (63%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5593	0.3474	0.1954	0.0308	0.6858	0.9304	0.4238
Bagging RPART	0.6207	0.2526	NaN	0.0000	0.6646	0.9743	NaN
RandomForest	0.6078	0.3240	0.3333	0.0093	0.6806	0.9619	0.4126
J48	0.5189	0.4095	0.2324	0.0599	0.6904	0.8817	0.4425
GBM	0.5979	0.4018	0.2432	0.0163	0.6922	0.9466	0.4369
C5.0	0.4902	0.3965	0.1071	0.0163	0.6944	0.9042	0.4174
JRip	0.5538	0.3070	NaN	0.0000	0.6734	0.9576	NaN
Nnet	0.4426	0.5070	NaN	0.0000	0.7058	0.8791	NaN
SVM	0.5063	0.3526	0.2474	0.0435	0.6883	0.9147	0.4251
NB	0.5000	0.0053	0.4444	0.0072	0.6304	0.9958	0.2656
OLR	0.5538	0.3070	NaN	0.0000	0.6719	0.9555	NaN

Table A.8: Results on J_25

J_25_Midden	High (16%)		Medium (17%)		Low (66%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.4956	0.2947	0.1628	0.0120	0.7017	0.9415	0.3987
Bagging RPART	0.5951	0.2140	NaN	0.0000	0.6872	0.9761	NaN
RandomForest	0.6371	0.2673	0.3590	0.0236	0.6967	0.9689	0.4105
J48	0.4826	0.2912	0.2917	0.0479	0.7002	0.9216	0.4138
GBM	0.5229	0.3404	0.3426	0.0634	0.7129	0.9260	0.4416
C5.0	0.5588	0.3000	0.3429	0.0205	0.7011	0.9535	0.4124
JRip	0.5427	0.2790	0.2903	0.0154	0.6982	0.9548	0.4014
Nnet	0.4883	0.3298	NaN	0.0000	0.7007	0.9393	NaN
SVM	0.5064	0.2772	0.1644	0.0205	0.6954	0.9322	0.3791
NB	0.0000	0.0000	NaN	0.0000	0.662	1.0000	NaN
OLR	0.5238	0.2316	NaN	0.0000	0.6899	0.9655	NaN

Table A.9: Results on J_25_Midden

J_25_Hoog	High (10%)		Medium (20%)		Low (70%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.3831	0.2362	0.2677	0.2224	0.7290	0.8039	0.4333
Bagging RPART	0.7111	0.0982	NaN	0.0000	0.7046	0.9958	NaN
RandomForest	0.7385	0.1399	0.2143	0.0045	0.7138	0.9937	0.3583
J48	0.3826	0.1845	0.2556	0.0518	0.7134	0.9322	0.3811
GBM	0.4658	0.2086	0.3362	0.1119	0.7260	0.9239	0.4231
C5.0	0.5238	0.1687	0.2895	0.0158	0.7117	0.9763	0.3695
JRip	0.5789	0.1350	NaN	0.0000	0.7075	0.9907	NaN
Nnet	0.3763	0.2239	0.2738	0.0330	0.7207	0.9476	0.3861
SVM	0.4632	0.1350	0.3380	0.0344	0.7124	0.9704	0.3644
NB	NaN	0.0000	0.0909	0.0014	0.6978	0.9962	NaN
OLR	0.5341	0.1442	0.0000	0.0000	0.7074	0.9861	NaN

Table A.10: Results on J_25_Hoog

J_50	High (35%)		Medium (14%)		Low (50%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5514	0.6238	NaN	0.0000	0.6421	0.7760	NaN
Bagging RPART	0.5521	0.6210	NaN	0.0000	0.6412	0.7780	NaN
RandomForest	0.5860	0.5794	0.2778	0.0111	0.6341	0.8229	0.4401
J48	0.5034	0.6310	0.1818	0.0400	0.6457	0.6817	0.4296
GBM	0.5852	0.6134	0.5000	0.0089	0.6488	0.8189	0.4468
C5.0	0.5942	0.5784	0.1429	0.0044	0.6327	0.8296	0.4376
JRip	0.6183	0.5312	NaN	0.0000	0.6211	0.8706	NaN
Nnet	0.5462	0.6040	NaN	0.0000	0.6326	0.7760	NaN
SVM	0.5416	0.5784	0.5000	0.0044	0.6263	0.7834	0.4214
NB	0.6059	0.1758	0.0000	0.0000	0.5331	0.9618	NaN
OLR	0.5374	0.5359	NaN	0.0000	0.6137	0.8001	NaN

Table A.11: Results on J_50

J_50_Laag	High (35%)		Medium (14%)		Low (51%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.6047	0.5528	0.1852	0.0112	0.6395	0.8417	0.4418
Bagging RPART	0.5674	0.5387	NaN	0.0000	0.6264	0.8201	NaN
RandomForest	0.6144	0.5783	0.1818	0.0095	0.6471	0.8334	0.4475
J48	0.4869	0.5761	0.0899	0.0179	0.6208	0.6829	0.4027
GBM	0.5938	0.6032	0.1094	0.0156	0.6598	0.8090	0.4509
C5.0	0.6004	0.5920	0.2727	0.0268	0.6508	0.8189	0.4567
JRip	0.5898	0.5275	NaN	0.0000	0.6250	0.8410	NaN
Nnet	0.5569	0.5892	NaN	0.0000	0.6376	0.7893	NaN
SVM	0.5651	0.5387	0.0000	0.0000	0.6272	0.8170	NaN
NB	0.5575	0.5023	0.0000	0.0000	0.6118	0.8195	NaN
OLR	0.5602	0.5210	NaN	0.0000	0.6198	0.8195	NaN

Table A.12: Results on J_50_Laag

J_50_Midden	High (35%)		Medium (14%)		Low (51%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5700	0.5429	0.1429	0.0043	0.6317	0.8259	0.4268
Bagging RPART	0.5863	0.5400	NaN	0.0000	0.6323	0.8451	NaN
RandomForest	0.5938	0.5457	0.2308	0.0130	0.6366	0.8414	0.4394
J48	0.4789	0.5524	0.1565	0.0498	0.6146	0.6731	0.4104
GBM	0.5998	0.5752	0.3077	0.0087	0.6487	0.8457	0.4461
C5.0	0.5856	0.5571	0.1579	0.0065	0.6367	0.8309	0.4348
JRip	0.6027	0.5057	NaN	0.0000	0.6175	0.8580	NaN
Nnet	0.5125	0.5838	NaN	0.0000	0.6250	0.7469	NaN
SVM	0.5471	0.4924	0.0000	0.0000	0.6078	0.8198	NaN
NB	0.7143	0.0095	0.2500	0.0022	0.5189	0.9975	0.2353
OLR	0.5475	0.4886	NaN	0.0000	0.6050	0.8198	NaN

Table A.13: Results on J_50_Midden

J_25-50	High (27%)		Medium (16%)		Low (57%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5206	0.3677	0.2533	0.0370	0.6311	0.8762	0.4098
Bagging RPART	0.5447	0.3554	NaN	0.0000	0.6270	0.9134	NaN
RandomForest	0.6050	0.3516	0.0000	0.0000	0.6238	0.9290	NaN
J48	0.5032	0.3654	0.1772	0.0275	0.6291	0.8632	0.3996
GBM	0.5541	0.4249	0.2881	0.0331	0.6457	0.8829	0.4288
C5.0	0.5659	0.3756	0.2222	0.0039	0.6305	0.9118	0.4015
JRip	0.5931	0.2713	NaN	0.0000	0.6109	0.9489	NaN
Nnet	0.4528	0.3980	0.2800	0.0136	0.6230	0.8291	0.3870
SVM	0.5153	0.3218	0.0909	0.0058	0.6129	0.8901	0.3777
NB	0.2500	0.0022	0.0000	0.0000	0.5627	0.9983	NaN
OLR	0.5443	0.2758	NaN	0.0000	0.6036	0.9229	NaN

Table A.14: Results on J_25-50

J_25-50_Laag	High (29%)		Medium (15%)		Low (56%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5202	0.4111	0.2093	0.0188	0.6298	0.8533	0.4061
Bagging RPART	0.5429	0.3515	NaN	0.0000	0.6192	0.9014	NaN
RandomForest	0.5841	0.4142	0.1667	0.0021	0.6326	0.8981	0.4104
J48	0.5006	0.4266	0.1290	0.0254	0.6344	0.8194	0.4061
GBM	0.5592	0.4615	0.1111	0.0021	0.6432	0.8715	0.4166
C5.0	0.5387	0.4296	0.1667	0.0021	0.6338	0.8693	0.4051
JRip	0.5659	0.3751	NaN	0.0000	0.6224	0.9009	NaN
Nnet	0.4767	0.4841	NaN	0.0000	0.6407	0.8056	NaN
SVM	0.5126	0.3967	0.1000	0.0042	0.6215	0.8555	0.3917
NB	0.5158	0.2353	0.0000	0.0000	0.5959	0.9286	NaN
OLR	0.5364	0.3556	NaN	0.0000	0.6106	0.8837	NaN

Table A.15: Results on J_25-50_Laag

J_25-50_Midden	High (26%)		Medium (15%)		Low (58%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.4607	0.2698	NaN	0.0000	0.6261	0.9128	NaN
Bagging RPART	0.4602	0.2660	NaN	0.0000	0.6254	0.9139	NaN
RandomForest	0.5112	0.2634	0.0000	0.0000	0.6250	0.9278	NaN
J48	0.4384	0.2909	0.1739	0.0343	0.6354	0.8679	0.3802
GBM	0.4945	0.3990	0.2553	0.0253	0.6476	0.8623	0.4091
C5.0	0.4415	0.3568	0.1778	0.0169	0.6456	0.8600	0.3877
JRip	0.4830	0.1995	NaN	0.0000	0.6129	0.9409	NaN
Nnet	0.4521	0.3261	NaN	0.0000	0.6275	0.8766	NaN
SVM	0.4330	0.2890	0.1333	0.0253	0.6225	0.8526	0.3696
NB	0.0000	0.0000	0.0000	0.0000	0.5815	0.9989	NaN
OLR	0.5194	0.2225	NaN	0.0000	0.6111	0.9340	NaN

Table A.16: Results on J_25-50_Midden

J_25-50_Hoog	High (26%)		Medium (18%)		Low (56%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.4705	0.3959	0.2326	0.0358	0.6395	0.8510	0.4074
Bagging RPART	0.5556	0.2906	NaN	0.0000	0.6204	0.9441	NaN
RandomForest	0.5617	0.3584	0.2000	0.0054	0.6362	0.9299	0.4012
J48	0.4735	0.4121	0.1761	0.0501	0.6442	0.8234	0.4138
GBM	0.5325	0.4467	0.3000	0.0268	0.6556	0.8861	0.4296
C5.0	0.5117	0.3705	1.0000	0.0018	0.6340	0.9069	0.3932
JRip	0.5499	0.3002	NaN	0.0000	0.6192	0.9359	NaN
Nnet	0.4731	0.4358	0.0857	0.0054	0.6410	0.8478	0.3979
SVM	0.4783	0.3036	0.1429	0.0107	0.6182	0.8938	0.3748
NB	0.2857	0.0024	0.0000	0.0000	0.5682	0.9967	NaN
OLR	0.5194	0.2760	NaN	0.0000	0.6115	0.9283	NaN

Table A.17: Results on J_25-50_Hoog

5. Results on subsamples without temporary work

All_Laag	High (42%)		Medium (19%)		Low (38%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5566	0.7175	0.2872	0.0377	0.5223	0.5943	0.4165
Bagging RPART	0.5384	0.6898	NaN	0.0000	0.5049	0.6145	NaN
RandomForest	0.5812	0.7182	0.3333	0.0126	0.5352	0.6682	0.4203
J48	0.5492	0.6818	0.2361	0.0711	0.5122	0.5638	0.4181
GBM	0.5895	0.7043	0.2500	0.0391	0.5409	0.6659	0.4354
C5.0	0.5427	0.6917	0.2254	0.0022	0.5232	0.6130	0.4045
JRip	0.5123	0.8277	NaN	0.0000	0.5698	0.4780	NaN
Nnet	0.5227	0.7835	NaN	0.0000	0.5384	0.5227	NaN
SVM	0.5556	0.7155	0.2353	0.0056	0.5209	0.6234	0.4013
NB	0.4261	0.9934	0.0000	0.0000	0.5897	0.0172	NaN
OLR	0.5508	0.7116	NaN	0.0000	0.5204	0.6271	NaN

Table A.18: Results on All_Laag

All_Midden	High (32%)		Medium (20%)		Low (48%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5333	0.5290	0.2353	0.0058	0.5662	0.7943	0.4012
Bagging RPART	0.5456	0.4683	NaN	0.0000	0.5546	0.8329	NaN
RandomForest	0.5773	0.5321	0.3750	0.0086	0.5691	0.8253	0.4148
J48	0.4987	0.5214	0.2234	0.0640	0.5703	0.7178	0.4149
GBM	0.5808	0.5792	0.2564	0.0144	0.5899	0.8177	0.4309
C5.0	0.5492	0.5550	0.2500	0.0130	0.5753	0.7931	0.4146
JRip	0.5955	0.4199	NaN	0.0000	0.5446	0.8734	NaN
Nnet	0.5172	0.5572	NaN	0.0000	0.5741	0.7784	NaN
SVM	0.5506	0.4982	0.3333	0.0072	0.5634	0.8236	0.4021
NB	0.3548	0.0097	0.0000	0.0000	0.4813	0.9883	NaN
OLR	0.5457	0.4727	NaN	0.0000	0.5590	0.8365	NaN

Table A.19: Results on All_Midden

All_Hoog	High 24%)		Medium (21%)		Low (55%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5009	0.3812	0.2807	0.0526	0.6324	0.8798	0.4191
Bagging RPART	0.4811	0.4290	NaN	0.0000	0.6318	0.8869	NaN
RandomForest	0.5706	0.4205	0.3330	0.0197	0.6376	0.9228	0.4252
J48	0.4496	0.4217	0.2690	0.0644	0.6435	0.8350	0.4220
GBM	0.5360	0.4501	0.2866	0.0740	0.6561	0.8745	0.4522
C5.0	0.5185	0.4346	0.1875	0.0099	0.6357	0.8945	0.4116
JRip	0.5684	0.3038	NaN	0.0000	0.6081	0.9446	NaN
Nnet	0.5192	0.3994	NaN	0.0000	0.6201	0.9022	NaN
SVM	0.5131	0.3868	0.2099	0.0280	0.6294	0.8898	0.4092
NB	0.2500	0.0014	0.5000	0.0016	0.5628	0.9982	0.2420
OLR	0.5345	0.4156	NaN	0.0000	0.6247	0.8851	NaN

Table A.20: Results on All_Hoog

J_25	High (24%)		Medium (21%)		Low (55%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5565	0.4203	0.3509	0.0554	0.6212	0.8848	0.4348
Bagging RPART	0.6471	0.2760	NaN	0.0000	0.5920	0.9609	NaN
RandomForest	0.6028	0.4812	0.3871	0.0166	0.6332	0.9186	0.4389
J48	0.5000	0.4451	0.2513	0.0653	0.6288	0.8359	0.4308
GBM	0.5780	0.5207	0.3228	0.0568	0.6461	0.8763	0.4628
C5.0	0.5889	0.4944	0.3889	0.0097	0.6307	0.908	0.4336
JRip	0.6180	0.4040	NaN	0.0000	0.6128	0.9360	NaN
Nnet	0.5748	0.4580	NaN	0.0000	0.6214	0.9117	NaN
SVM	0.5895	0.4793	0.3333	0.0139	0.6312	0.9117	0.4338
NB	0.3750	0.0075	0.2000	0.0028	0.5557	0.0994	0.2444
OLR	0.5692	0.4956	NaN	0.0000	0.6312	0.9064	NaN

Table A.21: Results on J_25

J_25_Midden	High (22%)		Medium (21%)		Low (58%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5519	0.3551	0.2562	0.0451	0.6207	0.8995	0.4144
Bagging RPART	0.5353	0.3011	NaN	0.0000	0.6061	0.9445	NaN
RandomForest	0.5861	0.3480	0.1818	0.0087	0.6182	0.9374	0.3995
J48	0.4384	0.3699	0.1613	0.0364	0.6160	0.8327	0.3896
GBM	0.5789	0.4276	0.3077	0.0698	0.6382	0.8891	0.4495
C5.0	0.5681	0.4091	0.2045	0.0131	0.6221	0.9094	0.4130
JRip	0.5721	0.3381	NaN	0.0000	0.6132	0.9418	NaN
Nnet	0.5395	0.3395	0.3108	0.0334	0.6172	0.9138	0.4046
SVM	0.5638	0.3452	0.2277	0.0334	0.6181	0.9099	0.4075
NB	0.6667	0.0028	NaN	0.0000	0.5673	1.0000	NaN
OLR	0.5544	0.3906	NaN	0.0000	0.6183	0.9226	NaN

Table A.22: Results on J_25_Midden

J_25_Hoog	High (10%)		Medium (21%)		Low (68%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.2992	0.2213	0.2324	0.1750	0.7111	0.7919	0.4011
Bagging RPART	0.6230	0.1064	NaN	0.0000	0.6969	0.9930	NaN
RandomForest	0.6418	0.1204	0.2500	0.0088	0.7006	0.9891	0.3467
J48	0.4057	0.1205	0.2742	0.0250	0.7006	0.9656	0.3479
GBM	0.4596	0.2073	0.3000	0.0971	0.7199	0.9256	0.4141
C5.0	0.5752	0.1821	0.2941	0.0221	0.7088	0.9782	0.3799
JRip	0.5603	0.1821	NaN	0.0000	0.7023	0.9839	NaN
Nnet	0.3774	0.2241	0.3516	0.0662	0.7161	0.9334	0.4010
SVM	0.4638	0.1793	0.3467	0.0382	0.7084	0.9626	0.3812
NB	0.0000	0.0000	0.0000	0.0000	0.6888	0.9983	NaN
OLR	0.4667	0.1569	NaN	0.0000	0.7038	0.9848	NaN

Table A.23: Results on J_25_Hoog

J_50	High (47%)		Medium (17%)		Low (36%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5980	0.7342	0.2973	0.0188	0.5189	0.6025	0.4174
Bagging RPART	0.6010	0.6892	NaN	0.0000	0.4942	0.6407	NaN
RandomForest	0.6042	0.7632	0.3333	0.0085	0.5435	0.6161	0.4229
J48	0.5396	0.7692	0.2282	0.0585	0.5015	0.4065	0.3921
GBM	0.6088	0.7471	0.2143	0.0103	0.5422	0.6340	0.4250
C5.0	0.6074	0.7568	0.2222	0.0137	0.5357	0.6100	0.4234
JRip	0.5333	0.8764	NaN	0.0000	0.5633	0.3693	NaN
Nnet	0.5873	0.7426	0.4444	0.0068	0.5234	0.5950	0.4087
SVM	0.5790	0.7523	0.3333	0.0068	0.5278	0.5751	0.4061
NB	0.5846	0.5959	0.1667	0.0017	0.4584	0.6672	0.3790
OLR	0.5567	0.7773	NaN	0.0000	0.5230	0.5095	NaN

Table A.24: Results on J_50

J_50_Laag	High (47%)		Medium (16%)		Low (36%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.6044	0.7488	0.1875	0.0116	0.5504	0.6136	0.4237
Bagging RPART	0.5507	0.7929	NaN	0.0000	0.5408	0.4787	NaN
RandomForest	0.6132	0.7803	0.2917	0.0136	0.5584	0.6009	0.4305
J48	0.5758	0.7423	0.1626	0.0390	0.5270	0.5118	0.4102
GBM	0.6132	0.7481	0.2571	0.0174	0.5536	0.6292	0.4319
C5.0	0.6041	0.7346	0.2500	0.0116	0.5472	0.6301	0.4237
JRip	0.5338	0.8635	0.3182	0.0136	0.5880	0.3751	0.3813
Nnet	0.5827	0.7631	0.0000	0.0000	0.5379	0.5622	NaN
SVM	0.5870	0.7739	0.5556	0.0097	0.5484	0.5666	0.4147
NB	0.4709	0.9790	0.1000	0.0019	0.5000	0.0287	0.2314
OLR	0.5764	0.7712	NaN	0.0000	0.5424	0.5509	NaN

Table A.25: Results on J_50_Laag

J_50_Midden	High (46%)		Medium (17%)		Low (37%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5724	0.7534	0.1852	0.0089	0.5008	0.5053	0.3902
Bagging RPART	0.5473	0.8187	NaN	0.0000	0.5148	0.4081	NaN
RandomForest	0.5996	0.7753	0.3182	0.0126	0.5384	0.5619	0.4168
J48	0.5330	0.7482	0.1923	0.0450	0.4910	0.3979	0.3783
GBM	0.6031	0.7393	0.2609	0.0216	0.5273	0.5870	0.4199
C5.0	0.6011	0.7589	0.0000	0.0000	0.5324	0.5862	NaN
JRip	0.5377	0.8871	NaN	0.0000	0.5942	0.3498	NaN
Nnet	0.5626	0.7576	NaN	0.0000	0.5045	0.4996	NaN
SVM	0.5730	0.7454	0.2500	0.0126	0.5087	0.5231	0.3959
NB	0.4791	0.9884	0.0000	0.0000	0.4255	0.0162	NaN
OLR	0.5584	0.7650	NaN	0.0000	0.5021	0.4818	NaN

Table A.26: Results on J_50_Midden

J_25-50	High (38%)		Medium (20%)		Low (43%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5266	0.5419	0.0000	0.0000	0.5129	0.7298	NaN
Bagging RPART	0.5606	0.4919	NaN	0.0000	0.5113	0.7994	NaN
RandomForest	0.5563	0.5886	0.2727	0.0046	0.5420	0.7579	0.4044
J48	0.5022	0.5498	0.2030	0.0637	0.5092	0.6266	0.3946
GBM	0.5626	0.6111	0.3846	0.0231	0.5489	0.7431	0.4203
C5.0	0.5305	0.5886	0.2759	0.0123	0.5339	0.7157	0.3977
JRip	0.5714	0.4541	NaN	0.0000	0.5006	0.8191	NaN
Nnet	0.5100	0.5982	NaN	0.0000	0.5288	0.6904	NaN
SVM	0.5235	0.5475	0.1765	0.0046	0.5215	0.7326	0.3845
NB	0.6475	0.0636	0.0000	0.0000	0.4374	0.9810	NaN
OLR	0.5306	0.5451	NaN	0.0000	0.5177	0.7417	NaN

Table A.27: Results on J_25-50

J_25-50_Laag	High (42%)		Medium (19%)		Low (39%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5395	0.7144	0.1333	0.0032	0.5310	0.5929	0.3938
Bagging RPART	0.5511	0.7101	NaN	0.0000	0.5220	0.6092	NaN
RandomForest	0.5722	0.7172	0.2857	0.0032	0.5330	0.6395	0.4081
J48	0.5442	0.7058	0.2955	0.0629	0.5169	0.5455	0.4164
GBM	0.5724	0.7265	0.2963	0.0129	0.5419	0.6325	0.4163
C5.0	0.5526	0.7372	0.1250	0.0048	0.5212	0.5726	0.3956
JRip	0.5097	0.8462	NaN	0.0000	0.5949	0.4530	NaN
Nnet	0.5388	0.7023	NaN	0.0000	0.5240	0.6030	NaN
SVM	0.5625	0.6823	0.4000	0.0065	0.5144	0.6387	0.3997
NB	0.6516	0.2664	0.0000	0.0000	0.4323	0.9184	NaN
OLR	0.5439	0.6765	NaN	0.0000	0.5062	0.6137	NaN

Table A.28: Results on J_25-50_Laag

J_25-50_Midden	High (37%)		Medium (19%)		Low (43%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.5032	0.5028	0.2791	0.0185	0.5015	0.7135	0.3756
Bagging RPART	0.5218	0.4761	NaN	0.0000	0.4982	0.7608	NaN
RandomForest	0.5408	0.5615	0.3889	0.0110	0.5360	0.7493	0.3991
J48	0.4900	0.5667	0.1794	0.0516	0.5103	0.6047	0.3864
GBM	0.5363	0.5815	0.1934	0.0092	0.5305	0.7204	0.3955
C5.0	0.5297	0.6002	0.0769	0.0015	0.5306	0.7051	0.3904
JRip	0.5667	0.3788	NaN	0.0000	0.4870	0.8456	NaN
Nnet	0.4789	0.6067	NaN	0.0000	0.5230	0.6398	NaN
SVM	0.5273	0.5020	0.5000	0.0031	0.5119	0.7629	0.3777
NB	0.6161	0.1054	0.2500	0.0031	0.4468	0.9638	0.2656
OLR	0.5033	0.4972	NaN	0.0000	0.5007	0.7323	NaN

Table A.29: Results on J_25-50_Midden

J_25-50_Hoog	High (32%)		Medium (21%)		Low (48%)		Overall
Technique	PPV	Recall	PPV	Recall	PPV	Recall	MacroF
RPART	0.4905	0.5369	0.2873	0.0362	0.5708	0.7472	0.4080
Bagging RPART	0.5980	0.2306	NaN	0.0000	0.5187	0.9454	NaN
RandomForest	0.5453	0.4783	0.1081	0.0058	0.5642	0.8356	0.3980
J48	0.4714	0.4629	0.2051	0.0465	0.5638	0.7545	0.3961
GBM	0.5458	0.5236	0.1852	0.0145	0.5762	0.8166	0.4123
C5.0	0.5077	0.4698	0.2115	0.0159	0.5643	0.8129	0.3946
JRip	0.5701	0.3497	NaN	0.0000	0.5341	0.8945	NaN
Nnet	0.4937	0.5539	NaN	0.0000	0.5748	0.7730	NaN
SVM	0.5251	0.4546	0.2963	0.0116	0.5571	0.8325	0.3924
NB	0.2727	0.0057	0.4286	0.0043	0.4833	0.9932	0.2233
OLR	0.5400	0.4216	NaN	0.0000	0.5523	0.8650	NaN

Table A.30: Results on J_25-50_Hoog

6. Results of different time intervals

Interval	Records (H/L)	Technique	Precision	Recall	MacroF
All					
3 months	6628 x 2841 test: 41/59% train: 41/59%	J48	0.5784	0.4562	0.6134
		GBM	0.6353	0.5103	0.6559
		RF	0.6658	0.4296	0.6396
		JRip	0.5789	0.4948	0.6239
		NB	0.7843	0.0344	0.4059
		Nnet	0.5965	0.5043	0.6355
6 months	6628 x 2841 test: 24/76% train: 25/75%	J48	0.5311	0.2954	0.6170
		GBM	0.6158	0.3372	0.6521
		RF	0.6778	0.2637	0.6265
		JRip	0.5556	0.3242	0.6339
		NB	0.1667	0.0014	0.4314
		Nnet	0.5360	0.2147	0.5817
12 months	6628 x 2841 test: 12/88% train: 12/88%	J48	0.4820	0.1914	0.6028
		GBM	0.5227	0.1971	0.6101
		RF	0.6744	0.0829	0.5421
		JRip	0.4933	0.2114	0.6140
		NB	0.0000	0.0000	NaN
		Nnet	0.3018	0.1457	0.5579
All (- 89)					
3 months	7225 x 3097 test: 54/46% train: 52/48%	J48	0.6480	0.6235	0.6135
		GBM	0.6817	0.7062	0.6610
		RF	0.6850	0.7026	0.6630
		JRip	0.6591	0.6930	0.6376
		NB	0.7826	0.0108	0.3269
		Nnet	0.6664	0.6001	0.6227
6 months	7225 x 3097 test: 34/66% train: 32/68%	J48	0.5202	0.3813	0.6037
		GBM	0.6107	0.4023	0.6419
		RF	0.6275	0.3051	0.6061
		JRip	0.5833	0.3136	0.5997
		NB	0.8000	0.0038	0.4020
		Nnet	0.5426	0.3823	0.6124
12 months	7225 x 3097 test: 8/92% train: 16/84%	J48	0.4537	0.0902	0.5254
		GBM	0.5502	0.2118	0.6050
		RF	0.6719	0.0792	0.5242
		JRip	0.5517	0.1179	0.5491
		NB	0.7500	0.0055	0.4576
		Nnet	0.3970	0.1952	0.5753

Table A.31: Results of time intervals

7. Results of Cox PH model

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
CATWZ_INSTROOM2	1.846e-01	1.203e+00	4.127e-02	4.473	7.72e-06	***
CATWZ_INSTROOM3	-1.020e-01	9.030e-01	3.838e-02	-2.657	0.007887	**
CATWZ_INSTROOM5	-4.999e-01	6.066e-01	9.172e-02	-5.450	5.02e-08	***
CATWZ_INSTROOM6	6.944e-02	1.072e+00	5.807e-01	0.120	0.904814	
CATWZ_INSTROOM11	-6.089e-01	5.440e-01	1.148e-01	-5.303	1.14e-07	***
CATWZ_INSTROOM14	-5.242e-01	5.920e-01	1.158e-01	-4.526	6.02e-06	***
LEEFTIJD_START_WKLSHD	-1.641e-02	9.837e-01	1.566e-03	-10.485	< 2e-16	***
PROVINCIELimburg	1.542e-02	1.016e+00	3.729e-02	0.413	0.679267	
PROVINCIENANA	4.628e-01	1.589e+00	3.042e-01	1.521	0.128184	
PROVINCIEOost-Vlaanderen	6.102e-02	1.063e+00	3.283e-02	1.859	0.063081	.
PROVINCIEVlaams-Brabant	-9.081e-03	9.910e-01	4.134e-02	-0.220	0.826138	
PROVINCIEwest-Vlaanderen	1.502e-01	1.162e+00	3.602e-02	4.168	3.07e-05	***
DGRAGMN	3.129e-01	1.367e+00	5.139e-02	6.089	1.14e-09	***
DGRALLN	1.281e-01	1.137e+00	3.451e-02	3.711	0.000206	***
EIGEN_WAGENJ	4.018e-01	1.495e+00	1.137e-01	3.533	0.000411	***
EIGEN_WAGENN	2.190e-01	1.245e+00	1.117e-01	1.961	0.049935	*
EIGEN_WAGENNANA	-1.903e-01	8.267e-01	3.251e-01	-0.585	0.558376	
RIJBEWIJSA3	3.588e-01	1.432e+00	1.944e-01	1.845	0.064987	.
RIJBEWIJSB	1.723e-01	1.188e+00	1.613e-01	1.068	0.285447	
RIJBEWIJSBE	1.952e-01	1.216e+00	1.768e-01	1.104	0.269566	
RIJBEWIJSC	2.528e-01	1.288e+00	1.827e-01	1.384	0.166496	
RIJBEWIJSC1	4.297e-01	1.537e+00	2.688e-01	1.598	0.109959	
RIJBEWIJSCE	4.415e-01	1.555e+00	1.884e-01	2.343	0.019135	*
RIJBEWIJSD	3.153e-01	1.371e+00	2.246e-01	1.404	0.160432	
RIJBEWIJSD1	-8.171e-03	9.919e-01	4.769e-01	-0.017	0.986331	
RIJBEWIJSDE	6.248e-01	1.868e+00	2.529e-01	2.470	0.013494	*
RIJBEWIJSG	1.562e-01	1.169e+00	2.467e-01	0.633	0.526551	
RIJBEWIJSNANA	1.429e-01	1.154e+00	1.603e-01	0.892	0.372640	
CAT_ARBGESCH1	1.343e-01	1.144e+00	1.087e-01	1.236	0.216417	
CAT_ARBGESCH2	-5.567e-03	9.944e-01	1.652e-01	-0.034	0.973117	
GESLACHTV	-3.550e-02	9.651e-01	2.439e-02	-1.456	0.145425	
STUDIE_NIVEAULaaggeschoold	-1.689e-01	8.446e-01	3.728e-02	-4.530	5.91e-06	***
STUDIE_NIVEAUMiddengeschoold	-7.481e-02	9.279e-01	3.228e-02	-2.317	0.020486	*
STUDIE_NIVEAUNANA	1.248e-02	1.013e+00	2.018e-01	0.062	0.950692	
TAAL_N1	-1.566e-01	8.550e-01	1.149e-01	-1.363	0.172972	
TAAL_N2	-4.752e-02	9.536e-01	1.110e-01	-0.428	0.668535	
TAAL_N3	-1.707e-02	9.831e-01	1.112e-01	-0.154	0.877937	
TAAL_NNANA	9.836e-01	2.674e+00	1.692e-01	5.815	6.07e-09	***
TAAL_F1	3.876e-01	1.473e+00	2.814e-01	1.378	0.168355	
TAAL_F2	3.545e-01	1.425e+00	2.818e-01	1.258	0.208494	
TAAL_F3	3.036e-01	1.355e+00	2.832e-01	1.072	0.283770	
TAAL_FNANA	4.670e-01	1.595e+00	2.817e-01	1.658	0.097344	.
TAAL_E1	-1.830e-01	8.328e-01	4.557e-01	-0.402	0.687980	
TAAL_E2	-1.829e-01	8.329e-01	4.554e-01	-0.402	0.688001	
TAAL_E3	-2.244e-01	7.990e-01	4.562e-01	-0.492	0.622693	
TAAL_ENANA	-7.538e-02	9.274e-01	4.558e-01	-0.165	0.868644	
WRKLS_PERIODES_VOOR_INSTR	5.136e-02	1.053e+00	4.091e-03	12.555	< 2e-16	***
DAGEN_WRKLS_10J_VOOR	-9.981e-04	9.990e-01	5.371e-05	-18.582	< 2e-16	***
CVS_INSTROOM	1.154e-02	1.012e+00	1.786e-02	0.646	0.518071	
GEWENSTE_JOBS_INSTROOM	-6.915e-02	9.332e-01	1.466e-01	-0.472	0.637144	
GEWENSTE_JOBS_ERV_1_INSTROOM	6.758e-02	1.070e+00	1.468e-01	0.460	0.645344	
GEWENSTE_JOBS_ERV_2_INSTROOM	1.011e-01	1.106e+00	1.468e-01	0.689	0.491090	
GEWENSTE_JOBS_ERV_3_INSTROOM	4.881e-02	1.050e+00	1.469e-01	0.332	0.739649	
GEWENSTE_JOBS_ERV_4_INSTROOM	1.307e-01	1.140e+00	1.471e-01	0.889	0.374139	
GEWENSTE_KNELPUNTBEROEPEN	-3.368e-03	9.966e-01	1.222e-02	-0.276	0.782898	
GEWENSTE_UITZONDERING_BEROEPEN	-3.690e-02	9.638e-01	2.024e-02	-1.823	0.068316	.
MAX_RATIO_VAC_PER_KANDIDATEN	1.195e-02	1.012e+00	1.436e-02	0.832	0.405328	
OPL_UREN_VOOR_INSTROOM	2.370e-05	1.000e+00	3.223e-05	0.736	0.462019	
AANTAL_BEDRIJVEN	8.399e-03	1.008e+00	1.587e-03	5.292	1.21e-07	***
ARBEID_DAGEN	-4.888e-07	1.000e+00	7.336e-06	-0.067	0.946876	
AANTAL_ARBEIDSCONTRACTEN	6.110e-04	1.001e+00	1.329e-04	4.597	4.28e-06	***
ARBEID_DAGEN_INTERIM	4.153e-04	1.000e+00	6.612e-05	6.281	3.35e-10	***
CONTINENTAZIE	-1.846e-01	8.314e-01	1.052e-01	-1.754	0.079346	.
CONTINENTBE	-1.121e-01	8.940e-01	6.966e-02	-1.609	0.107633	
CONTINENTNAM	-2.667e-01	7.659e-01	2.766e-01	-0.964	0.335002	
CONTINENTOTH	-1.505e-01	8.603e-01	7.645e-02	-1.969	0.048999	*
CONTINENTZAM	3.745e-01	1.454e+00	1.857e-01	2.016	0.043757	*
INTERESSES_INSTROOM_DISCTRUE	-1.737e-01	8.405e-01	5.367e-02	-3.237	0.001209	**
REFERS_INSTROOM_DISCTRUE	5.168e-02	1.053e+00	4.067e-02	1.271	0.203892	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure A.1: Results of Cox Regression

List of Figures

1.1 Time to event instances 3

4.1 KDD process by Fayyad et al (1996)[44] 8

5.1 Survival curve of both samples 16

6.1 Classifiers over all samples (left), without temporary work (right) 21
6.2 Classifiers over all samples (left), without temporary work (right) 22
6.3 Top three performing classifiers 22
6.4 Top three performing classifiers (-89) 23
6.5 Random forest compared to support vector machines on all samples 24
6.6 GBM and RF compared to SVM 24
6.7 Comparison of different time intervals 25
6.8 Comparison of different time intervals in sample without temporary work 25

A.1 Results of Cox Regression 44

List of Tables

1.1 Unemployment data by VDAB (December 2015) 2

3.1 Overview of related works 7

4.1 Example of a record 9
4.2 Removals in case 1 9
4.3 Removals in case 2 10
4.4 Removals in case 3 10
4.5 Class labels 11
4.6 Different samples used to train models 14

5.1 Significant attributes 17
5.2 Results from general sample with temporary work 17
5.3 Results from general sample without temporary work 18
5.4 Results from sample J_25_Laag 18
5.5 Results from sample J_50_Hoog 19
5.6 Results from sample J_25_Laag (-89) 20
5.7 Results from sample J_50_Hoog (-89) 20

7.1 Results based on Low & High class label 26

A.1 List of all attributes 32
A.2 Catwz_in codes 32
A.3 Catwz_uit codes 33
A.4 Inappropriate codes 33
A.5 Results on All_Laag 33
A.6 Results on All_Midden 34
A.7 Results on All_Hoog 34
A.8 Results on J_25 34
A.9 Results on J_25_Midden 35
A.10 Results on J_25_Hoog 35
A.11 Results on J_50 35
A.12 Results on J_50_Laag 36
A.13 Results on J_50_Midden 36
A.14 Results on J_25-50 36
A.15 Results on J_25-50_Laag 37
A.16 Results on J_25-50_Midden 37
A.17 Results on J_25-50_Hoog 37

A.18 Results on All_Laag 38
A.19 Results on All_Midden 38
A.20 Results on All_Hoog 39
A.21 Results on J_25 39
A.22 Results on J_25_Midden 39
A.23 Results on J_25_Hoog 40
A.24 Results on J_50 40
A.25 Results on J_50_Laag 40
A.26 Results on J_50_Midden 41
A.27 Results on J_25-50 41
A.28 Results on J_25-50_Laag 41
A.29 Results on J_25-50_Midden 42
A.30 Results on J_25-50_Hoog 42
A.31 Results of time intervals 43

Bibliography

Articles

[2] VDAB-Studiedienst, "Vdab werkloosheidsbericht december 2015", 2015, `https://www.vdab.be/trendsdoc/berichten/Werkloosheidsbericht_december_2015.pdf`.

[6] R. Boey, "Landurige werkloosheid in vlaanderen", *Over.werk Tijdschrift Van Het Steunpunt WSE*, no. 4, pp. 24–31, 2015.

[7] V. Ciuca and M. Matei, "Survival analysis for the unemployment duration", in *Proceedings of the 5th WSEAS International Conference on Economy and Management Transformation*, vol. 1, 2010, pp. 354–359.

[8] VDAB, "Paper5: Onderzoek naar de langetermijndynamiek van de werkloosheid", 2011.

[9] L. Desmet, "Een onderzoek naar de determinanten van de uitstroom naar werk, directie statistieken en studies rva", 2010.

[12] J. Lu, "Predicting customer churn in the telecommunications industry—-an application of survival analysis modeling using sas", *SAS User Group International (SUGI27) Online Proceedings*, pp. 114–27, 2002.

[13] J. Lu and O Park, "Modeling customer lifetime value using survival analysisan application in the telecommunications industry", *Data Mining Techniques*, pp. 120–128, 2003.

[14] L. Fu and H. Wang, "Estimating insurance attrition using survival analysis", *Variance-journal*, vol. 8, no. 1, 2014.

[15] W. C. Levy, D. Mozaffarian, D. T. Linker, S. C. Sutradhar, S. D. Anker, A. B. Cropp, I. Anand, A. Maggioni, P. Burton, M. D. Sullivan, *et al.*, "The seattle heart failure model prediction of survival in heart failure", *Circulation*, vol. 113, no. 11, pp. 1424–1433, 2006.

[16] F. Leonardo, J. M. Jerez, and E. Alba, "Artificial neural networks and prognosis in medicine. survival analysis in breast cancer patients", in *Proceedings of the 13th European Symposium on Artificial Neural Networks*, 2005, pp. 91–102.

[17] J Llobera, M Esteva, J Rifa, E Benito, J Terrasa, C Rojas, O Pons, G Catalan, and A Avella, "Terminal cancer: Duration and prediction of survival time", *European Journal of Cancer*, vol. 36, no. 16, pp. 2036–2043, 2000.

[18] A. Barth, L. A. Wanek, and D. L. Morton, "Prognostic factors in 1,521 melanoma patients with distant metastases.", *Journal of the American College of Surgeons*, vol. 181, no. 3, pp. 193–201, 1995.

[19] P. C. Adams, M. Speechley, and A. E. Kertesz, "Long-term survival analysis in hereditary hemochromatosis.", *Gastroenterology*, vol. 101, no. 2, pp. 368–372, 1991.

[20] C Rozman, E Montserrat, J. Rodriguez-Fernandez, R Ayats, T Vallespi, R Parody, A Rios, D Prados, M Morey, and F Gomis, "Bone marrow histologic pattern–the best single prognostic parameter in chronic lymphocytic leukemia: A multivariate survival analysis of 329 cases", *Blood*, vol. 64, no. 3, pp. 642–648, 1984.

[21] J. Binet, A Auquier, G Dighiero, C. Chastang, H Piguet, J Goasguen, G Vaugier, G Potron, P Colona, F Oberling, *et al.*, "A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis", *Cancer*, vol. 48, no. 1, pp. 198–206, 1981.

[22] S. S. Anand, A. E. Smith, P. Hamilton, J. Anand, J. G. Hughes, and P. Bartels, "An evaluation of intelligent prognostic systems for colorectal cancer", *Artificial Intelligence in Medicine*, vol. 15, no. 2, pp. 193–214, 1999.

[23] L. Bottaci, P. J. Drew, J. E. Hartley, M. B. Hadfield, R. Farouk, P. W. Lee, I. M. Macintyre, G. S. Duthie, and J. R. Monson, "Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions", *The Lancet*, vol. 350, no. 9076, pp. 469–472, 1997.

[24] H. B. Burke, "Artificial neural networks for cancer research: Outcome prediction", in *Seminars in Surgical Oncology*, Wiley Online Library, vol. 10, 1994, pp. 73–79.

[25] H. B. Burke, P. H. Goodman, D. B. Rosen, D. E. Henson, J. N. Weinstein, F. E. Harrell, J. R. Marks, D. P. Winchester, and D. G. Bostwick, "Artificial neural networks improve the accuracy of cancer survival prediction", *Cancer*, vol. 79, no. 4, pp. 857–862, 1997.

[26] L. Ohno-Machado, "A comparison of cox proportional hazards and artificial neural network models for medical prognosis", *Computers in biology and medicine*, vol. 27, no. 1, pp. 55–65, 1997.

[27] A. Bellaachia and E. Guven, "Predicting breast cancer survivability using data mining techniques", *Age*, vol. 58, no. 13, pp. 10–110, 2006.

[28] M. De Laurentiis and P. M. Ravdin, "A technique for using neural network analysis to perform survival analysis of censored data", *Cancer Letters*, vol. 77, no. 2, pp. 127–138, 1994.

[29] B. Zupan, J. DemšAr, M. W. Kattan, J. R. Beck, and I. Bratko, "Machine learning for survival analysis: A case study on recurrence of prostate cancer", *Artificial intelligence in medicine*, vol. 20, no. 1, pp. 59–75, 2000.

[30] W. N. Street, "A neural network model for prognostic prediction.", in *ICML*, Citeseer, 1998, pp. 540–546.

[31] C.-L. Chi, W. N. Street, and W. H. Wolberg, "Application of artificial neural network-based survival analysis on two breast cancer datasets", in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2007, 2007, p. 130.

[32] B. Baesens, T. Van Gestel, M. Stepanova, D. Van den Poel, and J. Vanthienen, "Neural network survival analysis for personal loan data", *Journal of the Operational Research Society*, vol. 56, no. 9, pp. 1089–1098, 2005.

[33] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes", *Journal of clinical epidemiology*, vol. 49, no. 11, pp. 1225–1231, 1996.

[34] A. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis", *Neural Computing and Applications*, vol. 23, no. 7-8, pp. 2387–2403, 2013.

[36] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: Bagging and random forests for ecological prediction", *Ecosystems*, vol. 9, no. 2, pp. 181–199, 2006.

[37] B. P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu, and G. McGregor, "Boosted decision trees as an alternative to artificial neural networks for particle identification", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 543, no. 2, pp. 577–584, 2005.

[38] Y. Chan, "Biostatistics 305. multinomial logistic regression", *Singapore medical journal*, vol. 46, no. 6, p. 259, 2005.

[39] J. Anderson, "Logistic regression", *Handbook of Statistics. North-Holland, New York*, pp. 169–191, 1982.

[40] B. Kempen, D. J. Brus, G. B. Heuvelink, and J. J. Stoorvogel, "Updating the 1: 50,000 dutch soil map using legacy soil data: A multinomial logistic regression approach", *Geoderma*, vol. 151, no. 3, pp. 311–326, 2009.

[41] Y. Wang, "A multinomial logistic regression modeling approach for anomaly intrusion detection", *Computers & Security*, vol. 24, no. 8, pp. 662–674, 2005.

[42] B. K. Bhardwaj and S. Pal, "Data mining: A prediction for performance improvement using classification", *ArXiv preprint arXiv:1201.3418*, 2012.

[43] Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks: A market comparative study", *Decision support systems*, vol. 37, no. 4, pp. 543–558, 2004.

[44] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *et al.*, "Knowledge discovery and data mining: Towards a unifying framework.", in *KDD*, vol. 96, 1996, pp. 82–88.

[47] J. Gholap, "Performance tuning of j48 algorithm for prediction of soil fertility", *Journal of Computer Science and Information Technology*, vol. 8, no. 2, 2012.

[48] S.-l. PANG and J.-z. GONG, "C5. 0 classification algorithm and application on individual credit evaluation of banks", *Systems Engineering-Theory & Practice*, vol. 29, no. 12, pp. 94–104, 2009.

[49] M. Kuhn, "Package 'caret'", 2016, `https://cran.r-project.org/web/packages/caret/caret.pdf`.

[50] T. Therneau, E. Atkinson, and M. Foundation, "An introduction to recursive partitioning using the rpart routines", 2015, `https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf`.

[51] L. Breiman, "Bagging predictors", *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[52] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: A classification and regression tool for compound classification and qsar modeling", *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[53] L. Breiman and A. Cutler, "Package 'randomforest'", 2015, `https://cran.r-project.org/web/packages/randomForest/randomForest.pdf`.

[54] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?", in *MLDM*, Springer, 2012, pp. 154–168.

[57] W. W. Cohen, "Fast effective rule induction", in *Proceedings of the twelfth international conference on machine learning*, 1995, pp. 115–123.

[58] W. Cohen, "Fast effective rule induction", 1995, http://www.csee.usf.edu/~hall/dm/ripper.pdf.

[60] R. Battiti and F. Masulli, "Bfgs optimization for faster and automated supervised learning", in *International neural network conference*, Springer, 1990, pp. 757–760.

[61] G.-B. Huang, Y.-Q. Chen, and H. A. Babri, "Classification ability of single hidden layer feedforward neural networks", *Neural Networks, IEEE Transactions on*, vol. 11, no. 3, pp. 799–801, 2000.

[67] T. M. Mitchell, "Machine learning", *Machine Learning*, 1997.

[68] M. F. Triola, "Bayes' theorem", *Elementary Statistics*, vol. 11, 2010.

[69] K. Ming Leung, "Naive bayes classifier", 2007, https://tom.host.cs.st-andrews.ac.uk/ID5059/L15-LeungSlides.pdf.

[70] Y. So and W. F. Kuhfeld, "Multinomial logit models", in *SUGI 20 Conference Proceedings*, 1995, pp. 1227–1234.

[72] B. Ripley, "Package 'mass'", 2016, https://cran.r-project.org/web/packages/MASS/MASS.pdf.

[74] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks", *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[76] Y. Yang and X. Liu, "A re-examination of text categorization methods", in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 1999, pp. 42–49.

[77] A. J. Viera, J. M. Garrett, *et al.*, "Understanding interobserver agreement: The kappa statistic", *Fam Med*, vol. 37, no. 5, pp. 360–363, 2005.

[78] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial", *Frontiers in neurorobotics*, vol. 7, 2013.

[79] R. Kanfer, C. R. Wanberg, and T. M. Kantrowitz, "Job search and employment: A personality–motivational analysis and meta-analytic review.", *Journal of Applied psychology*, vol. 86, no. 5, p. 837, 2001.

Books

[10] O. Chapelle, B. Schölkopf, A. Zien, *et al.*, *Semi-supervised learning*. MIT press Cambridge, 2006.

[35] G. Ilczuk and A. Wakulicz-Deja, "Attribute selection and rule generation techniques for medical diagnosis systems", in *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, Springer, 2005, pp. 352–361.

[46] S. Suthaharan, *Machine learning models and algorithms for big data classification: Thinking with examples for effective learning*. Springer, 2015, vol. 36, pp. 237–269.

[65] O. Maimon and L. Rokach, *Data mining and knowledge discovery handbook*. Springer, 2005, vol. 2.

[66] N. Cristianini and E. Ricci, "Support vector machines", in *Encyclopedia of Algorithms*, Springer, 2008, pp. 928–932.

[75] A. Özgür, L. Özgür, and T. Güngör, "Text categorization with class-based and corpus-based keyword selection", in *Computer and Information Sciences-ISCIS 2005*, Springer, 2005, pp. 606–615.

Internet

[1] https://www.vdab.be/vdab/algemeen.shtml.

[3] http://bestat.economie.fgov.be/BeStat.

[4] http://www.onprvp.fgov.be/NL/profes/benefits/retirement/age/paginas/default.aspx.

[5] http://www.belgium.be/nl/werk/.

[11] http://machinelearningmastery.com/an-introduction-to-feature-selection/.

[45] http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/.

[55] https://www.quora.com/Whats-the-difference-between-boosting-and-bagging.

[56] http://topepo.github.io/caret/Boosting.html.

[59] https://en.wikipedia.org/wiki/bfgs.

[62] http://topepo.github.io/caret/Neural_Network.html.

[63] http://stats.stackexchange.com/questions/29130/difference-between-neural-net-weight-decay-and-learning-rate.

[64] http://sebastianraschka.com/Articles/2014_about_feature_scaling.html.

[71] http://data.princeton.edu/wws509/r/c6s5.html.

[73] http://topepo.github.io/caret/splitting.html.

FACULTY OF BUSINESS AND ECONOMICS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be