A Case Study of Evaluating Job Readiness with Data Mining Tools and CRISP-DM Methodology

Izabela Wowczko
Institute of Technology Blanchardstown, Ireland

Abstract

This paper analyses the use of data mining techniques in evaluating job readiness of unemployed population in Ireland. To effectively help a jobseeker to enter or return to employment, it is necessary to develop a personal plan and provide them with suitable services. This report investigates how employment and further education needs are recognized among the customers of the Irish public employment services. Following the steps of CRISP-DM methodology, it explores the characteristics of the group and attempts to identify the underlying pattern. Finally, after applying suitable mining techniques, it is discussed whether the classification system with regard to job readiness can be automated.

1. Introduction

One of the principal governmental services offered to the public in any developed country is in the area of employment and further education. Three leading authorities cooperate to successfully match the expectations of employers with the abilities of the workforce on the Irish labour market. A national network of Intreo employment offices provides expert guidance and access to job vacancies. Further Education and Training Authority, SOLAS, is responsible for identifying skills shortages and targeting labour market demands by promoting relevant training programs. Finally, Education and Training Boards Ireland deliver and manage a wide range of further education services to bridge the gap between employers and jobseekers.

The following study focuses on a selected aspect of this complex subject, namely identification of support needs among customers registered with public employment services. Through the analysis of information acquired in a process of the registration, the general characteristics of the group are outlined.

Furthermore, the relationship between various futures in the database of customers is evaluated. Finally, the project investigates the existence of a pattern with regard to job readiness.

This study implements the standard data mining methodology. Information discovered at various stages of Cross Industry Standard Process for Data Mining (CRISP-DM) is a foundation for future recommendations regarding the predictive potential of the dataset. Software applied in this project is open source tools such as RapidMiner and Tableau, and commercial database management system SQL Server 2008.

The reminder of the paper is structured as follows: Section 2 provides an overview of CRISP-DM methodology; Sections 3-7 report on the execution of the consecutive steps, including the final conclusions.

2. CRISP-DM

CRISP-DM is a freely available model that has become the leading methodology in data mining. Because of its industry and tool independence, CRISP-DM can provide guidelines for organized and transparent execution of any project. Typically, it groups all scheduled tasks into six consecutive phases [1]:

- Business Understanding (understanding business objectives and converting them into data mining problems, developing project plan),
- Data Understanding (collection of data, verification of data quality, data exploration),
- Data Preparation (data selection, cleaning, data formatting and integration, constructing the final dataset for modelling),
- Modelling (selection of modelling techniques, building and assessment of a model),
- Evaluation (evaluation of results in the context of business objectives, reviewing the quality of the process, further decisions on deployment),
- Deployment (deployment plan, deployment execution and monitoring, final report and review of the project).

Despite clear distinction between the groups of tasks in CRISP-DM, it has been acknowledged that there is no single universal practice in information discovery. The variety of data and related issues as well as business objectives may require various degrees of flexibility in applying the CRISP-DM reference model. In practice, this usually refers to stepping back to earlier phases and improving the quality of a dataset to match the requirements of selected modelling techniques.

In evaluating job readiness, the biggest challenge proved to be the quality of the original dataset. It was therefore necessary, to intertwine data understanding and preparation phases. Moreover, the business objectives of the project did not require deployment of a model, thus the project was completed at the evaluation stage.

3. Step 1 – Business Understanding

Job readiness is one of the basic characteristics of a customer identified in the process of registration. Currently, it is evaluated by a case officer, based on the data provided in a registration form and obtained from a client during an initial interview. In the result, hundreds of officers are making independent judgments [2], deciding whether a client is ready to work (is a jobseeker) or requires further training (is a prospective trainee).

3.1. Business Objectives

To provide services matching the needs of a customer, they need to be correctly identified. Therefore, the business goal of this project is to analyse the registered unemployed population and examine the relationships between its various features captured in the database. Special attention is dedicated to job readiness, as it is the key determiner in the type of support offered to a client - job matching or further training opportunities.

3.2. Data Mining Objectives

CRISP-DM is the methodology applied to achieve business objectives of this assignment. Firstly, analysis of attributes and their subsets is conducted. Secondly, appropriate data mining tools are used in order to identify the underlying patterns.

4. Step 2 – Data Preparation

4.1. Dataset

The data used in this project is an extract from a real word dataset. It contains records of clients seeking guidance and support from the public employment services in Ireland. Due to privacy issues, the view of the original database was created, excluding all sensitive information such as first name, surname, Personal Public Service number and two address lines. A sample was derived for a 12 months registration period (year 2013) that includes 139 attributes and 60775 rows (prior to any pre-processing). This group of examples was considered to be representative of all unemployed people registered with public employment services in Ireland within a full year cycle.

Due to complexity of the dataset, the initial analysis was carried on with RapidMiner and SQL Server 2008. The Meta Data view (RapidMiner) and selection queries (SQL Server 2008) allowed comprehensive data quality assessment. In the first review of the data with the domain expert, it was agreed that the obsolete attributes and attributes with extremely high number of missing values can be discarded. 69 attributes were identified to hold no information value, and therefore excluded from further processing. The remaining 70 attributes (10 integer attributes, 56 nominal attributes and 4 text attributes) were closely analysed and pre-processed as explained further.

4.2. Data Cleaning and Pre-Processing

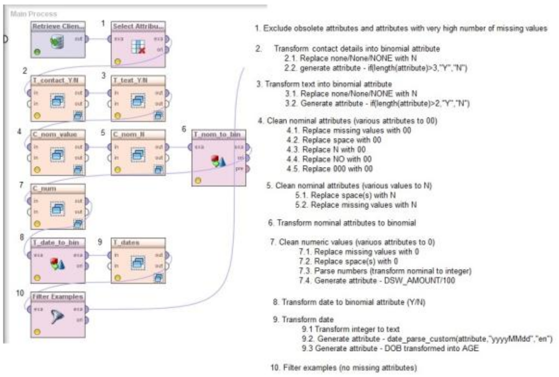
More detailed examination resulted in identifying a large number of missing values and incorrect values such as zeros, multiple zeros, ‘none’, ‘no’, spaces, multiple spaces, NULLs, etc. Among 70 attributes, only 8 did not require any data cleaning. It was also acknowledged that the majority of the attributes were of the wrong type. This resulted from the dirt in the data (i.e. binomial or integer attributes were regarded as nominal) or the generic type derived from the warehouse (i.e. dates were stored as integers). Therefore, it was decided that the data must be cleaned and transformed before carrying on any valid analysis. An in depth pre-processing plan was prepared in collaboration with the domain expert.

It was assumed that the missing and noisy values stand for ‘no’, ‘zero’, ‘did not happen’, etc. This simplification was reasonably justified and necessary to make the data manageable. The inconsistency in recorded values resulted from the lack of constraints on the data entered manually by a front desk staff in multiple offices throughout the country.

It was also deducted that, although values of some attributes might not be relevant in their raw form, they contain information that might be useful. The general idea behind that approach was the hypothesis that the more information clients provide the higher their job readiness might be. It was therefore agreed, that the best solution to retaining the informational value of those variables was to convert them into binomial attributes (Y, N). 7 attributes were cleaned and transformed with this method (GENERAL_COMMENTS, EMAIL_ADDRESS, MOBILE_NUMBER, PHONE_NUMBER, WORK_SKILLS, COMPUTER_PACKAGES and SPECIAL_NEEDS_REQS).

All dates originally identified as integers were also subject to conversion. It was observed, that some of them had missing values that could not have been calculated i.e. missing INERVIEW_DATE simply means that a client has not yet been interviewed. On the other hand, filtering those examples resulted in a loss of too much data. Those variables were, therefore, also converted into binomial attributes (Y, N). Moreover, a new AGE attribute was calculated and included in a dataset to replace DOB (date of birth).

The data was transformed into the final set with the use of RapidMiner as detailed in Figure 1. At the cleaning and pre-processing stage of the project each attribute was evaluated separately in terms of its quality and informational value. The process outputted a clean set, however it was acknowledged that majority of identified issues could be easily eliminated via remodelling the warehouse i.e. putting constrains on data inputted into tables.



1. Exclude obsolete attributes and attributes with very high number of missing values

2. Transform contact details into binomial attribute
   2.1. Replace none/None/NONE with N
   2.2. generate attribute - if(length(attribute)>3,"Y","N")

3. Transform text into binomial attribute
   3.1. Replace none/None/NONE with N
   3.2. Generate attribute - if(length(attribute)>2,"Y","N")

4. Clean nominal attributes (various attributes to 00)
   4.1. Replace missing values with 00
   4.2. Replace space with 00
   4.3. Replace N with 00
   4.4. Replace NO with 00
   4.5. Replace 000 with 00

5. Clean nominal attributes (various values to N)
   5.1. Replace space(s) with N
   5.2. Replace missing values with N

6. Transform nominal attributes to binomial

7. Clean numeric values (variuos attributes to 0)
   7.1. Replace missing values with 0
   7.2. Replace space(s) with 0
   7.3. Parse numbers (transform nominal to integer)
   7.4. Generate attribute - DSW_AMOUNT/100

8. Transform date to binomial attribute (Y/N)

9. Transform date
   9.1 Transform integer to text
   9.2. Generate attribute - date_parse_custom(attribute,"yyyyMMdd","en")
   9.3 Generate attribute - DOB transformed into AGE

10. Filter examples (no missing attributes)

Figure 1. Cleaning and pre-processing routine in RapidMiner (T-transformation technique, C-cleaning technique)

5. Step 3 – Data Understanding

The processes discussed in Section 3.2. increased the readability and interpretability of the sample. Further analysis was carried on in RapidMiner (exploratory statistical analysis) and Tableau (data visualisation).

5.1. Exploratory Statistics (RapidMiner)

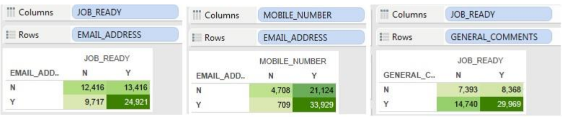
The majority of the attributes are binomial, therefore allow drawing generalised conclusions. However, a few numeric variables provide more insight into demographics of the customers registered in 2013. Some of the basic findings can be summarised as follows:

- Comparable numbers of women and men were registered in 2013 (48% and 52% respectively),
- The majority of customers had been interviewed (62%),
- The minority of customers require further training, whereas the majority are ready to enter employment (37% and 63% respectively),
- The average customer is about 35.5 years old, the youngest is 16 and the oldest is 101,
- Only little above 10% of the customers have a dependent adult,
- The maximum number of dependent children is 12, but the average value < 1 suggests that the vast majority of the customers are childless,
- Scatter plot of JOB_READY, AGE, and CHILD_DEP reveals that young people are generally ready to work if they are childless, but the more kids they have, the less job ready they are. As customers get older (and children grow up), they are more willing to return to employment,
- The vast majority of the registered clients receive financial aid,
- About 75% of the customers have at least 1 profession, whereas about 18% of them listed 3 profession they are experienced in,
- The maximum level of experience declared by clients is 2 (on a scale from 0 to 3),
- The majority of registered clients are poorly educated with no particular work skills.

5.2. Data Visualisation (Tableau)

Further trends and relationships between attributes were observed via data visualisations with Tableau. To make the analysis more informative, it was performed with the use of supplementary data: occupation names (to replace numeric MANCO codes) and geospatial coordinates (to visualise the geographic distribution of the examples). Some of the findings are:

- There is, in fact, a relationship between the amount of information provided by a client and their job readiness (Figure 2),
- The majority of clients are young, and the biggest number of customers registered in year 2013 was 20 years old, followed by 19 and 21,
- The majority of customers have some degree of professional experience. It can be also observer that the higher level of experience they declare, the more job ready they are. A vast number of clients with experience are willing to work, whereas clients with no experience are usually interested in further training,
- Relatively large percentage of the customers has not been paid any allowance. However, many have been paid for at least one year or have been on a long term financial support. In total, more payment was received by customers who are job ready,
- The biggest number of customers was registered in January, whereas the least number was registered in December.



Columns JOB_READY
Rows EMAIL_ADDRESS

EMAIL_ADD..	JOB_READY N	JOB_READY Y
N	12,416	13,416
Y	9,717	24,921

Columns MOBILE_NUMBER
Rows EMAIL_ADDRESS

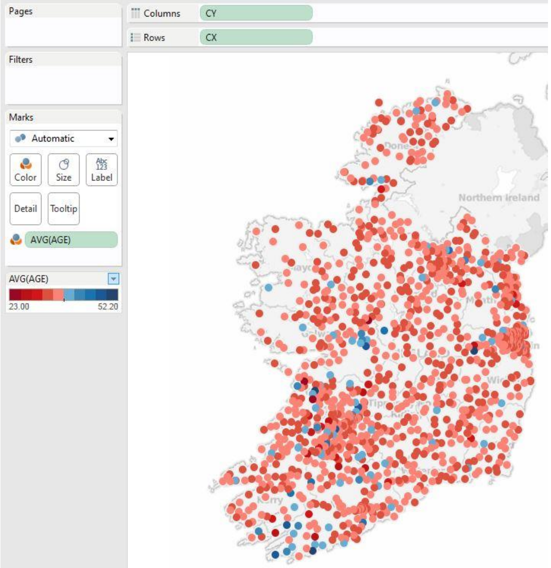
EMAIL_ADD..	MOBILE_NUMBER N	MOBILE_NUMBER Y
N	4,708	21,124
Y	709	33,929

Columns JOB_READY
Rows GENERAL_COMMENTS

GENERAL_C..	JOB_READY N	JOB_READY Y
N	7,393	8,368
Y	14,740	29,969

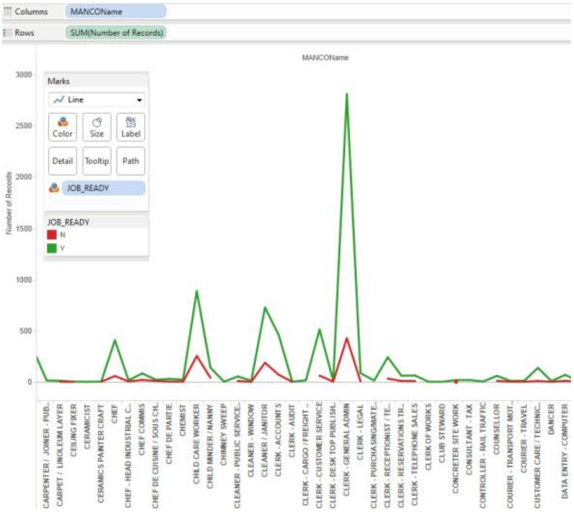
Figure 2. JOB_READY vs. information provided by a client

Supplementary data allowed performing more visually appealing analysis with regard to location, and the professional background among customers. Examples are pictured in Figure 3 and Figure 4.



Columns CY
Rows CX
Marks: Automatic
AVG(AGE)

Figure 3 Average customer age (Tableau)



Columns MANCOName
Rows SUM(Number of Records)
Marks: Line
JOB_READY

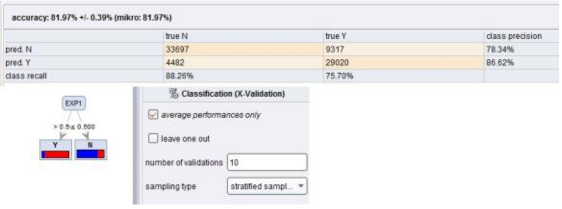
Figure 4 JOB_READY vs. professional background

6. Step 4 - Modelling

As previously pointed, the data mining objective of this project was not building a highly accurate model, but rather applying various techniques to investigate consistency in evaluating job readiness throughout the country. Therefore, it was decided to perform classification [3] with JOB_READY binomial class label (Y, N), using algorithms that provide the most informative feedback relevant to business objectives. Furthermore, the variables were analysed with correlation matrix, and various weighting techniques were applied to verify the link between job readiness and other attributes in the dataset. All processing was executed in RapidMiner.

6.1. Classification

As illustrated in Figure 5, classification with Decision Tree [4] achieved the accuracy of 81.97%. The algorithm built a very simple tree with just one split on attribute EXP1 (experience declared by a client with relation to their main profession - MANCO1). Model parameterisation (i.e. feature selection, normalisation) did not produce a bigger tree and did not improve its accuracy.



accuracy: 81.97% +/- 0.39% (mikro: 81.97%)

	true N	true Y	class precision
pred. N	33697	9317	78.34%
pred. Y	4482	29020	86.62%
class recall	88.26%	75.70%	

Classification (X-Validation)
☑ average performances only
□ leave one out
number of validations 10
sampling type stratified sampl...

Figure 5 Decision Tree (RapidMiner)

In a new approach, EXP1 attribute was excluded from the dataset to enable the algorithm finding a new split point. Another two variables were identified with this method: EXP2 **(**experience declared by a client with relation to their secondary profession – MANCO2) and FULL_TIME (willingness to work full-time).

A variation of the Decision Tree, a Random Forest [5], produced 20 trees with split points either on EXP2 or FULL_TIME. Although the achieved accuracy was much lower (57.61%), this result is consistent with previous findings.

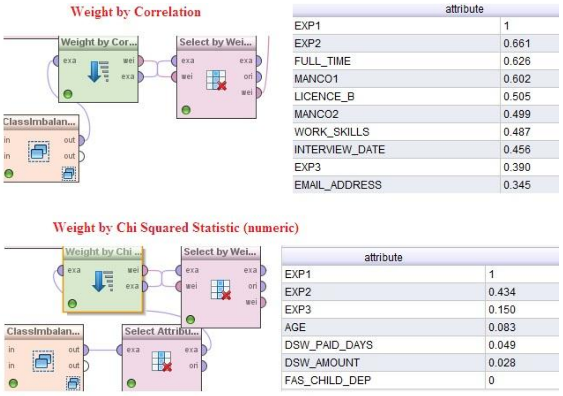
To further analyse the relationship between job readiness and other attributes, Support Vector Machine [6] was ran on the numeric attributes. Although the model achieved lower accuracy (58.64%) and mixed recall and precision for the two classes, the main attributes included in decision making - EXP1, EXP2, and EXP3 - were comparable with those identified earlier.

K-Nearest Neighbour [7] applied to the dataset achieved a relatively high accuracy of 75.15% and confirmed the existence of an underlying pattern. Unfortunately, this type of algorithm does not provide any useful information about the attributes.

6.2. Other Methods

Correlation matrix [8] was analysed to verify findings of the classification process. Although it identified close relationship between job readiness and MANCO1 and MANCO2, those variables are discarded by algorithms due to their high variability. Attributes such as EXP1 or FULL_TIME represented by small ranges of values have more predictive powers, and therefore are favoured by classification models.

The last method used in this experiment tested various weighting techniques [9]. As illustrated in Figure 6, the results are consistent with those of other techniques and identify the same attributes as being crucial in affecting job readiness.



attribute	
EXP1	1
EXP2	0.661
FULL_TIME	0.626
MANCO1	0.602
LICENCE_B	0.505
MANCO2	0.499
WORK_SKILLS	0.487
INTERVIEW_DATE	0.456
EXP3	0.390
EMAIL_ADDRESS	0.345

attribute	
EXP1	1
EXP2	0.434
EXP3	0.150
AGE	0.083
DSW_PAID_DAYS	0.049
DSW_AMOUNT	0.028
FAS_CHILD_DEP	0

Figure 6. Attributes weighting (RapidMiner)

7. Step 5 – Evaluation and Conclusion

Data mining techniques and CRISP-DM methodology confirmed the existence of a pattern with regard to job readiness among unemployed population registered with Irish public employment services. There is consistency in the examined sample, and the top discriminators between the two class labels are attributes such as EXP1, EXP2 and FULL_TIME. There is a very straightforward relationship among the features of the customers – a person is job ready if they have some work experience and are willing to be employed full-time. Otherwise, they decide on further training to improve their skills and gain relevant practice. This pattern, however, can be easily recognized and there was no previously unknown information discovered in the process of mining the dataset. It is also worth noticing that, in some cases, despite being fully qualified and experienced, a client might request training in order to change their career path (therefore not be JOB_READY).

Considering the findings of the project and acknowledging the specific of the domain, it has been concluded that automating the system would not offer any advantage over the existing classification based on simple heuristics [10]. However, it is apparent that increasing job readiness among clients can be achieved by delivering educational programs that include work placement, therefore provide an opportunity to gain initial work experience.

References

[1] Shearer, C. (2000), 'The CRISP-DM Model: The New Blueprint for Data Mining', *Journal of Data Warehousing, Volume 5, Number*, pp. 13-22

[2] Wang, J. & Liu, D. & Ruhe, G. (2004), 'Formal Description of the Cognitive Process of Decision Making', *Proceedings of the Third IEEE International Conference on Cognitive Informatics (ICCI'04)*, pp.124-130,
DOI: 10.1109/COGINF.2004.1327467

[3] Beniwal, S. & Arora, J. (2012) 'Classification and Feature Selection Techniques in Data Mining', *International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 6*, August – 2012, ISSN: 2278-0181

[4] Quinlan, J.R. (1986), 'Induction of Decision Trees', *Machine Learning Volume 1, Issue 1*, pp.81-106

[5] Breiman, L. (2001), 'Random Forest', Machine Learning October 2001, Volume 45, Issue 1, pp 5-32

[6] Shrivastava, D.K. & Bhambhu, L. (2010), 'Data Classification Using Support Vector Machine', *Journal of Theoretical and Applied Information Technology Vol 12. No. 1*; http://www.jatit.org/volumes/research-papers/Vol12No1/1Vol12No1.pdf (30 October 2014)

[7] Guo, G. & Wang, H. & Bell, D. & Bi, Y. & Greer, K. (2003), 'KNN Model-Based Approach in Classification', *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE Lecture Notes in Computer Science Volume 2888*, pp**.** 986-996

[8] Tiwari, R. & Singh, M.P. (2010), 'Correlation-based Attribute Selection using Genetic Algorithm', *International Journal of Computer Applications (0975 – 8887) Volume 4– No.8, August 2010*, pp. 28-34

[9] Song, Y.C. &, Meng, H.D. &, O'Grady, M.J. & O'Hare, G.M.P. (2007), 'Applications of Attributes Weighting in Data Mining', *Proceedings of the IEEE SMC UK&RI 6th Conference on Cybernetic Systems, Dublin, Ireland, September 2007*, pp. 41-45

[10] Hutchinson, J.M.C & Gigerenzer, G. (2005), 'Simple heuristics and rules of thumb: Where psychologists and behavioural biologists might meet', *Behavioural Processes 69 (2005,)* pp. 97–124