International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-5 Issue-6, August 2016

Analysis of Irish Labour Market using Predictive Modelling

A. Nachev

Abstract : This study explores empirically Irish labour market and factors affecting employability rate of Irish nationals, using data from the Quarterly National Household Survey and data mining techniques. The research is conducted according to the CRISP-DM methodology and addresses its stages. We perform data cleansing and reduction of dimensionality, analyse data, and build predictive models to measure employability rate. The study uses two statistical techniques to train the models and also provides performance analysis of the models, measures variable significance using sensitivity analysis (SA) and variable effect characteristic (VEC) curves. The paper discusses results and draws conclusions.

Index Terms: data mining, classification, logistic regression, linear discriminant analysis, labour market.

I. INTRODUCTION

Analyzing data from large or big-data sources, by the means of data mining techniques and methodologies, has become a valuable approach to discover patterns, relationships, and hidden information. The knowledge obtained in that way would serve decision making in various domains, including labour market management, employability management, HR management, and so on. Recently, great effort is made to analyse factors that determine the labour demand and supply in order to meet each other. Those who benefit from such analysis are employers, social and recruitment agencies, professional associations, educational institutions and bodies, and governments, all striving to create employment opportunities to the work force, and particularly to young people.

This study aims to analyse the Irish labour market specifics using nationwide data gathered by an Irish national survey conducted from 2014 to 2015. In general, knowledge discovery by data mining includes several methods, such as prediction/regression, classification, clustering, affinity analysis, etc. This study uses classification as one of the most prominent and effective supervised learning methods for building predictive models that fit the data available and then used for predictions and decision making.

Extant research in the domain discusses various aspects of labour and employability, mostly focusing to students and graduates as a target group [14], [15], [16] or to workers at organisational level for the purposes of HR management [17], [18]. Literature suggests, that the data mining techniques used for model building include decision trees and Bayesian methods [14], [15], [16], [17], [18], ensemble methods, MLP, and SVM [15]. The remainder of the paper is organized as follows: section II provides an overview of the CRISP-DM data mining methodology used; section III discusses the dataset used in the study, its features, the preprocessing steps needed to prepare the data for experiments, partitioning, and statistical analysis; section IV discusses the techniques used to build binary classifiers - logistic regression and linear discriminant analysis; section V presents and discuses experimental results; and section VI gives conclusions.

Manuscript published on 30 August 2016.
***** Correspondence Author (s)
A. Nachev, BIS, Cairnes Business School, NUI Galway, Galway, Ireland.

II. METHODOLOGY

Data mining as a process requires a number of different skills and knowledge. In order to be successful, the process needs a standard approach which would help translate business problems into data mining tasks, suggest appropriate data transformations and data mining techniques, and provide means for evaluating the effectiveness of the results. The CRISP-DM (CRoss Industry Standard Process for Data Mining) [2] is a standard framework for carrying out data mining projects, which is independent of both the industry sector and the technology used and widely adopted across researchers and practitioners in the field. This study adheres to the CRISP-DM methodology as research approach. Shortly, CRISP-DM consists of a cycle that comprises six stages (Figure 1): Business understanding stage focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and designs a preliminary plan to achieve the objectives. Data understanding stage starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, and discover first insights into the data. Data preparation stage covers all activities to construct the final dataset from the initial raw data. In the modeling stage various modeling techniques are selected and applied and their parameters are calibrated to optimal values. At the evaluation stage the model (or models) obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain that it properly achieves the business objectives. The deployment stage implements the model in a real environment [2]. With reference to the business understanding stage, our objective is to use the dataset to build a classification model, which predicts employability rate using empirical data. Formally, that is a binary classification task. Apart from the potential of applying such a model in practice, we also aim to analyse factors contributing to the Irish labor employability.

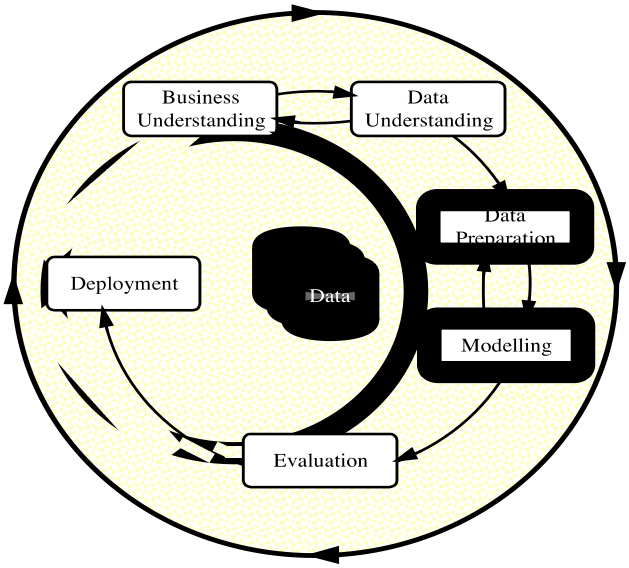


Figure 1. Phases of the CRISP-DM process model for data mining.

With reference to the data understanding stage, we process data as outlined below.

III. DATA

A. Dataset

This study uses a dataset obtained form the Irish Quarterly National Household Survey (QNHS) [1], which was conducted in 2014 and 2015. The information is collected continuously each week throughout the years using computer-assisted personal interview (CAPI) software. The survey data is presented on half-year basis in quarters Q2 and Q4 having number of records as follows: 2014 Q2 - 52,763; 2014 Q4 - 50,515; 2015 Q2 - 50,939; 2015 Q4 - 45,047 records. The original dataset contains 115 variables divided into 3 categories: 104 core variables; 7 derived variables for labour market analysis; and 4 derived variables for family unit analysis. The core variables are further split into the following groups: demographic background - 14 variables; labour status - 4; employment characteristics of main employment - 22; atypical work - 5; hours worked - 7; second job - 4; previous work experience of person not in employment - 8; search for employment - 3; method used to find work - 15; main labour status - 1; education and training - 9; dwelling unit information - 7; technical items relating to interviews - 5 variables.

B. Data Preprocessing

With reference to the data understanding and data preparation stages of the CRISP-DM standard for data mining projects [2], the dataset was cleansed by removing variables deemed as not relevant to the data mining goal. That includes 91 variables not relevant to all respondents; those not relevant to employment status; and not relevant to the period 2015-2016 covered. Further 8 variables were removed, as they were deemed dependent to other variables, form a business point of view. A new binary variable ILO_BIN was derived form the non-binary ILO (ILO/EU employment status of the respondent), in order to serve the binary classification task.

After the cleansing, the dataset contains 17 variables in five groups as follows:

Demographic: SEX (gender); MARSTAT (matial status); NATIONAL_SUMMARY(nationality of the respondent); YEARESID_SUMMARY (years of residence in this country).

Education: EDUCLEVL (education level); HATLEVEL (highest level of education successfully completed) HATFIELD (field of highest level of education successfully completed);

Dwelling unit information: DWELLINGUNIT (type of dwelling the respondent lives in); NUMBEROFROOMS (number of rooms); CONSTRUCTIONDATE (construction date of the dwelling); NATUREOFOCCUPANCY(nature of occupancy of the dwelling);

Technical items related to interview: REGION (region of household); AGECLASS (age class of the respondent);

Family status: FAMILYTYPE_SUMMARY (type of family); FAMILYPERSON_SUMMARY (person role within the family); FAMILYSTRUCTURE_SUMMARY (summary of family type)

Finally, the target variable is ILO_BIN.

With reference to the data preparation stage of CRISP-DM, the cleansing also involved discarding records of respondents whose age class is not relevant to the task - age below 16 or above 75. Also records containing missing values were removed.

The dataset was then broken into four subsets, each representing a quarter of the period covered: 2014 Q2 - 35978; 2014 Q4 - 30409; 2015 Q2 - 34240; 2015 Q4 - 28978 records, respectively.

C. Partitioning

Another pre-processing step is dataset partitioning, which prepares data for the modelling stage of CRISP-DM [2]. Typically, data have to be broken into two or three partitions: training, validation, and optionally testing. While the training partition is presented to the model to train and fit to its data, the validation partition is used by the training algorithm to tests the level of training until it reaches a satisfactory level. Despite the validation set is not directly involved in training, it indirectly influences the training by adapting the model hyper-parameters towards its own features and specifics. Thus, using a validation set only for the purpose of intermediate testing may result with merit figures, which are quite optimistic. Another robust and data neutral approach to produce realistic estimates is to set apart a testing partition, used solely for testing, once the model is built. Using such a dedicated test partition is particularly important when various modelling techniques are to be assessed and compared.

This study uses four dedicated test partitions for each of the subsets mentioned above. In fact, each subset was split into two partitions - one training & validation holding 80% of the data (split later on automatically at the modelling stage into separate training and validation in ratio 2:1), and one testing partition holding the rest of 20%. Data records for each partition are selected randomly.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

D. Data Summary Statistics

With reference to the CRISP-DM process, data understanding and preparation require computing of basic descriptive statistics. The range statistics (min, max, range) is useful for data checking to detect coding errors. Other summary statistics are also important for selection of modelling techniques and their design.

Summary statistics of the entire dataset is presented in Table 1. It is representative for each of the four quarters in 2014-2015, so separate analysis of these is not needed.

Table 1. Summary statistics of the QNHS dataset

	min	max	range	mean	sd	median	trimmed	mad	skew	kurtosis	se
sex	1	2	1	1.51	0.5	2	1.52	0	-0	-2	0
marstat	1	4	3	1.72	0.74	2	1.62	0	1.2	2	0
national_summary	1	6	5	1.27	0.95	1	1	0	3.7	13	0
yearesid_summary	0	13	13	1.28	3.32	0	0.29	0	2.4	4	0
educlevl	1	9	8	8.32	1.81	9	8.86	0	-3	5.1	0
hatlevel	0	800	800	362.56	184.1	304	358	154	0.3	-1	0.5
hatfield	0	9999	9999	3254.6	4487	500	2818	741	0.8	-1	12
dwellingunit	1	9	8	2.51	1.88	2	2.12	1.5	1.9	3.2	0
numberofrooms	1	9	8	3.85	1.61	4	3.79	1.5	0.5	0.4	0
constructiondate	1	12	11	6.98	3.39	7	7.12	4.5	-0	-1	0
natureofoccupancy	1	11	10	1.78	1.53	1	1.44	0	2.2	5.2	0
region	1	2	1	1.74	0.44	2	1.8	0	-1	-1	0
ageclass	4	15	11	9.18	3.24	9	9.16	4.5	0.1	-1	0
familytype_summary	1	9	8	2.97	2.6	2	2.46	0	1.8	1.4	0
familyperson_summary	1	9	8	2.9	2.67	2	2.38	1.5	1.7	1.2	0
familystructure_summary	1	9	8	5.29	2.68	5	5.36	4.5	-0	-1	0
ilo_bin	1	2	1	1.44	0.5	1	1.42	0	0.3	-2	0

In the table, *sd* is standard deviation, a quantity expressing by how much the members of a group differ from the mean value for the group. *Trimmed mean* is a statistical measure of central tendency, much like the mean and median. It involves the calculation of the mean after discarding given parts of a probability distribution or sample at the high and low end, and typically discarding an equal amount of both.

MAD is median absolute deviation, a robust measure for variability. Unlike the standard mean/standard deviation, MAD is not sensitive to presence of outliers.

Skew or skewness is a measure of the asymmetry of the probability distribution of a random variable about its mean. The skewness value can be positive or negative.

Kurtosis is a measure of the "tailedness" of the probability distribution of a random variable.

SE or standard error of the mean estimates the variability between sample means obtained if multiple samples from the same population are taken. The standard error of the mean estimates the variability between samples whereas the standard deviation measures the variability within a single sample of a population.

E. Correlation Analysis

Correlation analysis is used to assess the strength and direction of the linear relationships between pairs of variables. In statistical terms, correlation is a method of assessing a possible two-way linear association between two continuous variables. It is measured by correlation coefficients, which take values in the range −1 to +1. A correlation coefficient of zero indicates that no linear relationship exists between the variables, and a coefficient of −1 or +1 indicates a perfect linear relationship. The strength of relationship can be anywhere between −1 and +1. The stronger the correlation, the closer the correlation coefficient comes to ±1. If the coefficient is a positive number, the variables are directly related (i.e., as the value of one variable goes up, the value of the other also tends to do so). If, on the other hand, the coefficient is a negative number, the variables are inversely related (i.e., as the value of one variable goes up, the value of the other tends to go down). Any other form of relationship between two continuous variables that is not linear is not correlation in statistical terms. Figure 2 illustrates scatterplots and associated correlation coefficients.

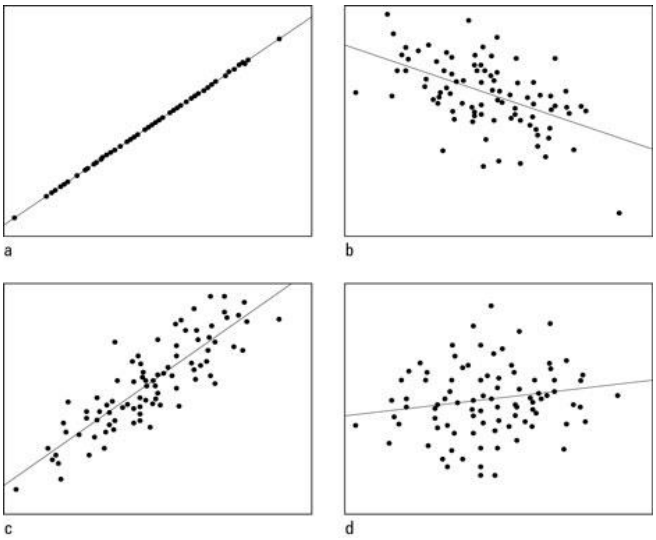


Figure 2. Scatterplots and corresponding correlation coefficients a) +1.00; b) –0.50; c) +0.85; and d) +0.15.

There are different methods for correlation analysis: Pearson parametric correlation test, Spearman and Kendall rank-based correlation tests. The default is Pearson correlation coefficient is computed by

$$r = \frac{\sum XY - \frac{1}{n}\sum X \sum Y}{\sqrt{(\sum X^2 - \frac{1}{n}(\sum X)^2)(\sum Y^2 - \frac{1}{n}(\sum Y)^2)}}, \quad (1)$$

where X and Y are two variables of size n.

Spearman's rank correlation coefficient is appropriate when one or both variables are skewed or ordinal and is robust when extreme values are present. Similar to Spearman's is the Kendall coefficient, which is appropriate for discrete variables. We used the Pearson's method to analyse QNHS variables, as it is suitable for the dataset we use.

In conclusion, the correlation analysis of QNHS is particularly useful for attribute relevance analysis carried out at the later modeling stage, when datasets undergo reduction of dimensionality in order to improve models. Along with other factors, a linear dependency would be a strong case for variable elimination. Figure 3 illustrates graphically correlation matrix of the QNHS dataset with 17 variables. The circle size and color represent coefficient values as per legend on the right.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

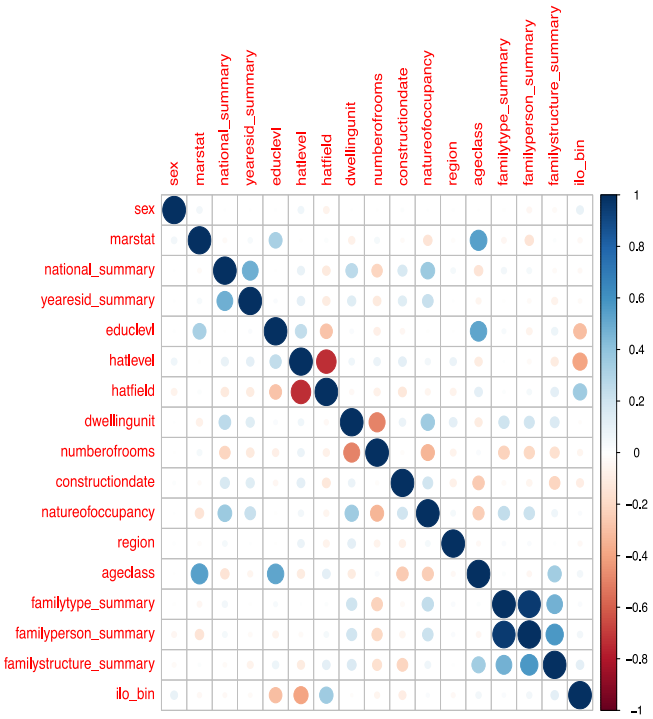


Figure 3. Pearson correlation matrix of QNHS, 17 variables.

Selecting variables with distinct dependencies (large circles) and grouping them together allows visualizing strongest dependencies, as sown in Figures 4-6.

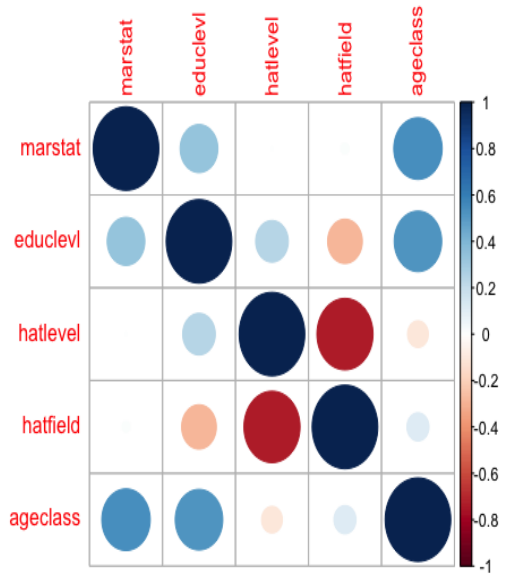


Figure 4. Correlations of group A variables.

Variable AGECLASS (age) shows 0.54 correlation with MARSTAT (marital status) and 0.53 with EDUCLEVEL (education level). This is explicable, as the education level naturally growth along with age, Similarly age goes along with marital status values 'single', 'married', 'divorced', and 'widowed'. Strong negative correlation of -0.73 exists between HATLEVEL (highest education completed) and HATFIELD (field of highest education completed). This can be explained by value distribution of those variables, where the lowest of one are related to the highest of the other at the two ends of the intervals.

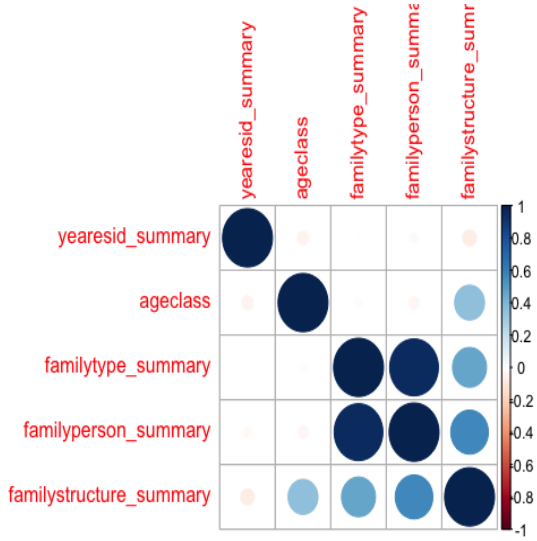


Figure 4. Correlations of group B variables.

A strong positive dependency of 0.95 can also be found between FAMILYPERSON_SUMMARY and FAMILYTYPE_SUMMARY, where naturally values like 'couple family unit' can be related to 'person is head' of a family, or 'partner'. Also, a dependency of 0.57 exists between FAMILYPERSON_SUMMARY and FAMILYSTRUCTURE_SUMMARY, explicable by the relations between values like 'family with children' and number of children, as well as 'lone parent family' and 'no children'.

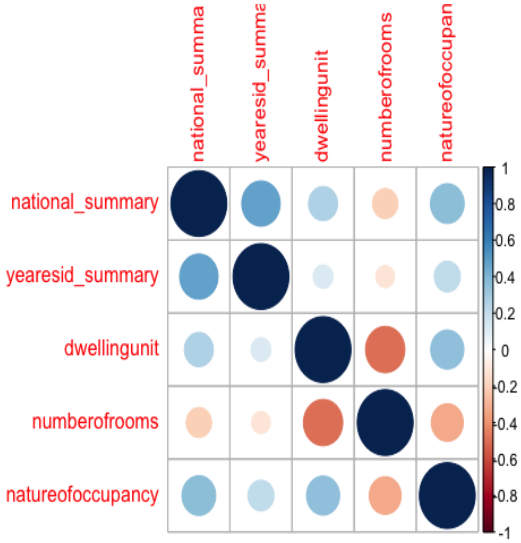


Figure 5. Correlations of group C variables.

Variables YEARESID (years of residence) and NATIONAL_SUMMARY (nationality of respondent) are positively related by 0.47, as for vast majority of the respondents, 'Irish' corresponds to 'born in the country'.

IV. CLASSIFICATION TECHNIQUES

With reference to the CRISP-DM modelling stage, this study considers two statistical techniques for building binary classifiers based on QNHS data from 2014 to 2015.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

These models would allow drawing conclusions about Irish labor market and explore factors, which affect employability rate.

A. Logistic Regression

Briefly, the logistic regression analysis is a statistical technique through which to examine relationship between a binary outcome - dependent variable and a set of predictors - independent variables. The outcome variable can be both continuous and categorical. If X_1 , X_2 ,..., X_n denote n predictor variables, Y denotes employed (Y = 1) or unemployed (Y = 0), and p denotes the probability of employment (i.e., the probability that $Y = 1$), the following equation describes the relationship between the predictor variables and p:

$$Logit(p) = \ln(\frac{p}{1- p}) = b_0 + b_1X_1 + ... + b_nX_n \quad (2)$$

where b_0 is a constant and b_1 , b_2 , ..., b_n are the regression coefficients of the predictor variables X_1, X_2, ..., X_n. The regression coefficients are estimated from the available data. The probability of employment p can be estimated with this equation.

Each regression coefficient describes the size of the contribution of the corresponding predictor variable to the outcome. The effect of the predictor variables on the outcome variable is commonly measured by using the odds ratio of the predictor variable, which represents the factor by which the odds of an outcome change for a one-unit change in the predictor variable. The odds ratio is estimated by taking the exponential of the coefficient. For example, if b_1 is the coefficient of variable AGECLASS ('age'), and p represents the probability of employment, exp(b_1) is the odds ratio corresponding to age class of the respondent, given all other predictor variables remain unchanged.

With this interpretation, we can describe logistic regression as a two-step procedure. First, we compute the logistic response function p

$$p = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2+\cdots+b_nx_n)}}, \quad (3)$$

which is a modification of (2) by solving for p. The next step is to use a cutoff value on these probabilities in order to map each case to one of the class labels. For example, in a binary case, a cutoff of 0.5 means that cases with an estimated probability of $p(Y=1) > 0.5$ are classified as belonging to class 1, whereas cases with $p(Y=1) < 0.5$ are classified as belonging to class 0. This cutoff need not be set at 0.5.

The regression coefficients b_i are usually estimated using maximum likelihood estimation [3]. Unlike linear regression with normally distributed residuals, it is not possible to find a closed-form expression for the coefficient values that maximize the likelihood function, so that an iterative process must be used instead, for example applying Newton's method. This process begins with a tentative solution, revises it slightly to see if it can be improved, and repeats this revision until improvement is minute, at which point the process is said to have converged [3]. In some instances the model may not reach convergence. Non-convergence of a model indicates that the coefficients are not meaningful because the iterative process is unable to find appropriate solutions. A failure to converge may occur for a number of reasons, e.g. having a large ratio of predictors to cases, multicollinearity, sparseness, or complete separation.

Logistic regression is one of the most commonly used tools for applied statistics and discrete data analysis. There are basically four reasons for this. First, it is a traditional technique. Secondly, the quantity log p/(1 − p) plays an important role in the analysis of contingency tables. Classification is a bit like having a contingency table with two columns (classes) and infinitely many rows (values of x). With a finite contingency table, we can estimate the log-odds for each row empirically, by just taking counts in the table. With infinitely many rows, we need some sort of interpolation scheme; logistic regression is linear interpolation for the log-odds. Thirdly, it's closely related to an exponential family of distributions, where the probability of some vector v is proportional to $\exp b_0 + \sum_{i=1}^{m} f_i(v)b_i$.

If one of the components of v is binary, and the functions f_i are all the identity function, then we get a logistic regression. Exponential families arise in many contexts in statistical theory and in physics, so there are lots of problems, which can be turned into logistic regression. Finally, the logistic regression often works surprisingly well as a classifier.

B. Linear Discriminant Analysis

Linear discriminant analysis (LDA), formulated by Fisher [4], is a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

LDA is closely related to analysis of variance (ANOVA) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements [5]. However, ANOVA uses categorical independent variables and a continuous dependent variable, whereas LDA has continuous independent variables and a categorical dependent variable, i.e. the class label.

LDA is also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables, which best explain the data. PCA can be described as an unsupervised algorithm, as it ignores class labels and its goal is to find the directions (the so-called principal components) that maximize the variance in a dataset. In contrast to PCA, LDA is supervised and computes the directions (linear discriminants) that will represent the axes that that maximize the separation between multiple classes.

Although it might sound intuitive that LDA is superior to PCA for a multi-class classification task where the class labels are known, this might not always the case.

Listed below are the 5 general steps for performing a linear discriminant analysis:

1. Compute the d-dimensional mean vectors for the different classes from the dataset.
2. Compute the scatter matrices (in-between-class and within-class scatter matrix). The within-class scatter matrix S_W is computed by

$$S_W = \sum_{i=1}^{c} S_i \text{ , where}$$

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$S_i = \sum_{x \in D_i}^{n} (x - m_i)(x - m_i)^T$ is scatter matrix for every class and m_i is the mean vector. The between-class scatter matrix S_B is computed by $S_B = \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T$, where m is the overall mean, and m_i and N_i are the sample mean and sizes of the respective classes.

3. Compute the eigenvectors (e_1, e_2, ... ,e_d) and corresponding eigenvalues (λ_1, λ_2, ... ,λ_d) for the scatter matrices by solving the generalized eigenvalue problem for the matrix $S_W^{-1} S_B$. Both eigenvectors and eigenvalues provide information about the distortion of a linear transformation. The eigenvectors are basically the direction of this distortion, and the eigenvalues are the scaling factor for the eigenvectors that describing the magnitude of the distortion.
4. Select linear discriminants for the new feature subspace via sorting the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a $d \times k$ - dimensional matrix W (where every column represents an eigenvector).
5. Use this $d \times k$ eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: $Y = X \times W$ (where X is a $n \times d$ - dimensional matrix representing the n samples, and Y are the transformed $n \times k$ - dimensional samples in the new subspace).

It should be mentioned that LDA assumes normal distributed data, features that are statistically independent, and identical covariance matrices for every class. However, LDA can also work reasonably well if those assumptions are violated [7].

V. MODELLING AND DISCUSSION

With reference to the CRISP-DM modelling stage, we trained and tested a number of predictive models based on the statistical techniques logistic regression (LR) and linear discriminant analysis (LDA). All experiments were carried out using R environment [8], [9]. The variable ILO_BIN was used as binary output, containing class labels 'employed' and 'unemployed' and all other variables were used as predictors. With each technique we did separate models for each of the QNHS subsets 2014 Q2, 2014 Q4, 2015 Q2, and 2015 Q4. These subsets were randomly split in ratio 2:1 into training and validation partitions. In order to validate results and reduce the effect of lucky set composition, each technique with a subset were tested many times, as follows: internally, the fit algorithm run 10 times using different random selection of training and validation sets. For each fit, we applied 3-fold cross-validation (CV) creating three model instances and then averaged their results. Finally, when models were built, they were tested using the 20% test partition, as discussed in section III *C*.

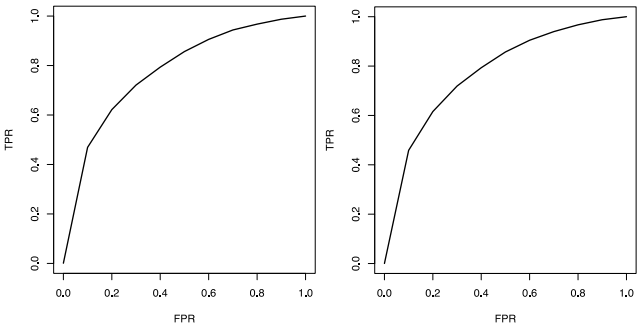
A. Results and Models Performance

In data mining, classification performance is often measured using accuracy (ACC) as the figure of merit. For a given operating point of a classifier, the accuracy is the total number of correctly classified instances divided by the total number of all available instances. Accuracy, however, varies dramatically depending on class prevalence. It can be a misleading estimator in cases where the most important class is underrepresented. In some cases sensitivity and specificity can be more relevant performance estimators. In order to address the accuracy deficiencies, we did Receiver Operating Characteristics (ROC) analysis [13]. In a ROC curve, the true positive rate (TPR), a.k.a. sensitivity, is plotted as a function of the false positive rate (FPR), a.k.a. 1-specificity, for different cut-off points. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold. A model with perfect discrimination between the two classes has a ROC plot that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC plot is to the upper left corner, the higher the overall accuracy of the model. The area under the ROC curve (AUC) is a common measure for the evaluation of discriminative power. AUC represents classifier performance over all possible threshold values, i.e. it is threshold independent.

Table 1 shows performance estimators for each modeling technique applied to each subset of data, all measured by ACC and AUC. It is evident that the two techniques build similar models with variances in accuracy ranging from 71.7% to 74.2% and AUC value ranging from 0.79 to 0.81. Figure 6 also illustrates that ROC curves of LR and LDA are nearly identical, no matter which period is considered.

Table 1. Logistic Regression (LR) and Linear Discriminant Analysis (LDA) performance measured by accuracy (ACC) and area under the ROC curve (AUC).

Metric	*'14 Q2*	*'14 Q4*	*'15 Q2*	*'15 Q4*
ACC_{LR}	72.1%	73.4%	74.2%	73.1%
AUC_{LR}	0.788	0.798	0.807	0.798
ACC_{LDA}	71.7%	73.2%	73.8%	72.9%
AUC_{LDA}	0.786	0.797	0.807	0.796



a) **b)**

Figure 6. ROC curve for a) Logistic Regression b) Linear Discriminant Analysis models.

B. Variable Significance

With refference to the data preparation and modelling stages of CRISP-DM [2], we can determine which variables significantly contribute to the model. There are many ways to estimate a variable's contribution to the model, and some may be better than others.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Estimation of the significance is important, as relatively few variables may appear explicitly as class discriminators. That means that other variables are not important in understanding or predicting the dependent variable. In the QNHS data context, the variable importance can be interpreted as role of factors, which influence employability rate of the respondents.

This work uses Sensitivity Analysis (SA) for ranking the variable importance to the model by measuring the effects on the output when the inputs are varied through their range of values [11]. While initially proposed for neural nets, SA is currently used with virtually any supervised learning technique. In summary, the SA varies an input variable x_a through its range with L levels, under a regular sequence from the minimum to the maximum value. Let $x_{a,j}$ denotes the j-th level of input x_a. Let $\hat{y}$ denote the value predicted by the model for one data sample (x) and let $\hat{y} = P(x)$ is the function of model responses. Kewley, Embrechts, and Breneman propose in [12] three sensitivity measures, namely range (S_r), gradient (S_g) and variance (S_v):

$$\begin{aligned} S_r &= \max(\hat{y}_{a_j} : j \in \{1,\ldots,L\}) - \min(\hat{y}_{a_j} : j \in \{1,\ldots,L\}) \\ S_g &= \sum_{j=2}^{L} \left| \hat{y}_{a_j} - \hat{y}_{a_{j-1}} \right| / (L-1) \\ S_v &= \sum_{j=2}^{L} (\hat{y}_{a_j} - \bar{y}_a)^2 / (L-1) \end{aligned} \quad (4)$$

where $\bar{y}_a$ denotes the mean of the responses. The gradient is the only measure that is dependent on the order of the sensitivity responses. For all measures, the higher the value, the more relevant is the input x_a. The relative importance r_a can be given by:

$$r_a = V_a / \sum_{i=1}^{M} V_i \quad (5)$$

where V_a is the sensitivity measure for x_a (e.g., range) [11]. Figures 7-9 show ranking of variable significance using the measures gradient, variance, and range averaging results form 10 runs, each with whiskers representing the variances in results. The figures illustrate that the variable education level (EDUCLEVEL) is top ranked, followed by the field of highest education completed (HATLEVEL) and age class of the respondent (AGECLASS), given the latter two take either second or third place depending on the measure used. These results reveal that according to the empirical study, the top three major factors that affect the Irish labour market and people's employability are related mostly to education and age of the people. This conclusion would drive actions that institutions at any level, from government down to social and recruitment agencies, professional associations, educational institutions and bodies individual, should take to address education, particularly targeting specific age groups. On the other end, the least significant variables for the models are region of residence, nationality of the respondents, and their family structure. Apparently, the Irish labour market shows homogeneity with respect to those features, so that they do not contribute to the models ability to discriminate between employed and unemployed classes.

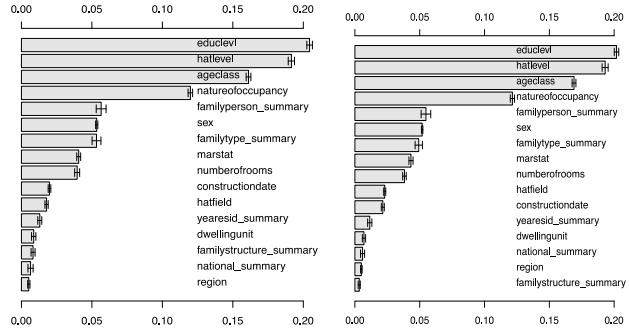


Figure 7. Variable significance for a) Logistic Regression b) Linear Discriminant Analysis models using 'gradient' sensitivity measure.

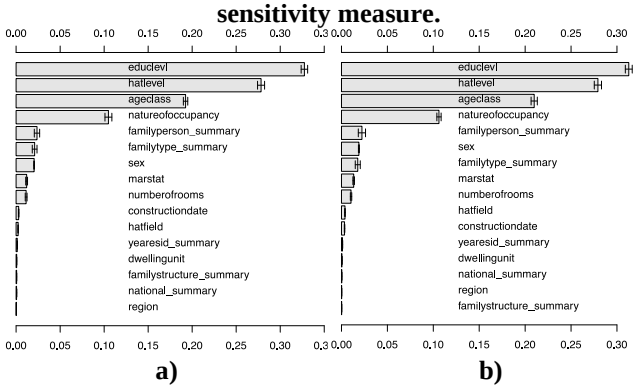


Figure 8. Variable significance for a) Logistic Regression b) Linear Discriminant Analysis models using 'variance' sensitivity measure.

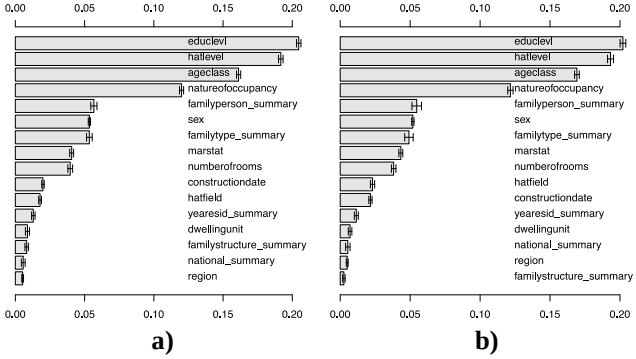


Figure 9. Variable significance for a) Logistic Regression b) Linear Discriminant Analysis models using 'range' sensitivity measure.

Finally, we built variable effect characteristic (VEC) curves [10] to explore the average impact of the three most significant variables x_a, which plot the x_{a_j} values (x-axis) versus the $\hat{y}_{a_j}$ responses (y-axis). Between two consecutive x_{a_j} values, the VEC plot uses a line (interpolation) for numerical values and a horizontal segment for categorical data. Figures 10 a) - c) show how EDUCLEVEL, HATLEVEL, and AGECLASS contribute the model performance.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

From the EDUCLEVEL (education level) VEC it is evident that education is not only important factor affecting the Irish labor market, but also it contributes exponentially to that. Max sensitivity of that variable is about 0.65. Similarly, HATLEVEL (highest education level completed) VEC shows that the higher the completed level is, the better for employability. That relation is close to linear, with level of max sensitivity 0.85. On the other hand, the AGECLASS (age) plays negative role to the employability, overall. That relation is nearly linear, reveling that with aging, chances for employment reduce.

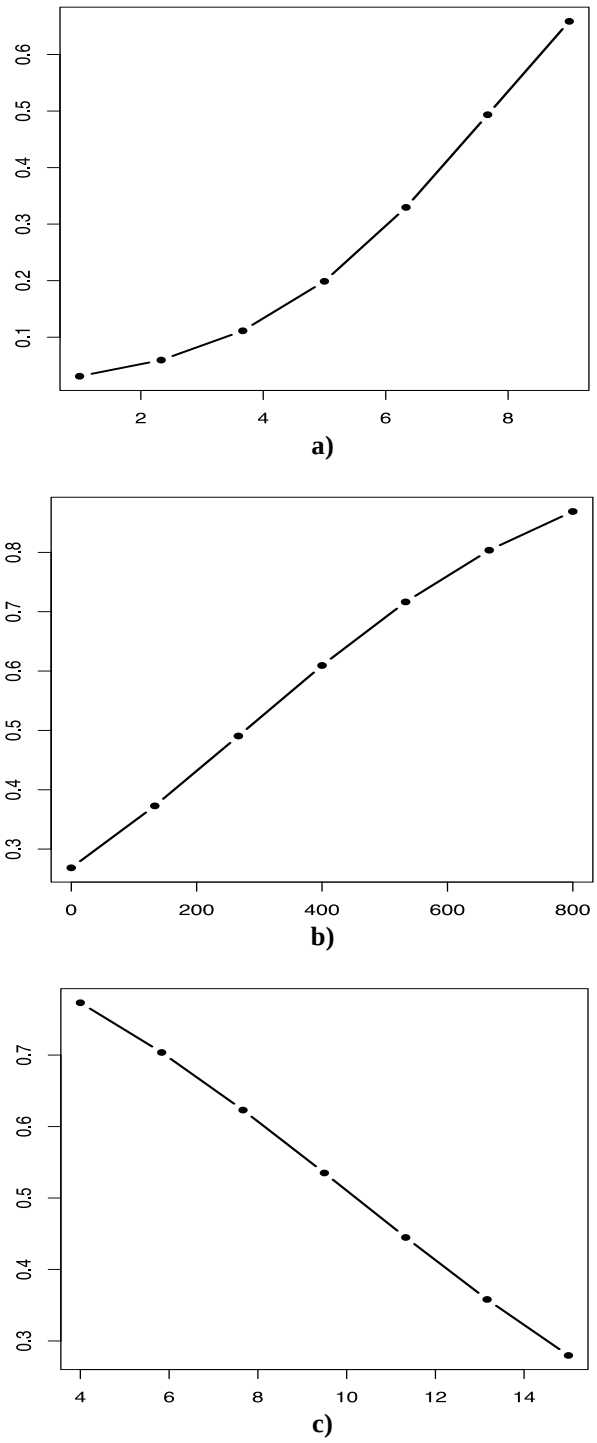


Figure 10. VEC curves of variables a) EDUCLEVEL, b) HATLEVEL, and c) AGECLASS.

VI. CONCLUSION

This paper presents an empirical analysis of the Irish labour market using data mining techniques and predictive modelling. Data source is the Irish Quarterly National Household Survey. The objective of this study is twofold: first, having a good fit to the data would allow applying the predictor as part of an advisory tool. Secondly, analysing the model, its variable significance, and variable effect characteristics allows drawing conclusions about the labour market itself and the factors affecting it.

Research methodology used is CRISP-DM and the paper adheres to the stages of that model. The objective was to build a binary classifier that fit the data and predicts employability rate. We did data pre-processing, discarding 98 out of 115 variables, those deemed to be irrelevant to the data mining task or dependent to other. We also removed samples irrelevant to the task and those containing missing values. As part of the pre-processing, we did correlation analysis of the data. In order to avoid bias in testing, we partitioned the dataset into training, validation, and testing sets. The testing partition (20%) was used for testing solely and wasn't presented to the models at the training stage.

Two statistical techniques were used for modelling, logistic regression and linear discriminant analysis. In order to provide a solid validation of results and avoid risk of 'lucky set' composition caused by the randomness, each model was built 10 times and results were averaged. Also, each run applied 3-fold cross-validation during the training.

Model performance was estimated by both accuracy and ROC analysis with AUC metric. All models showed consistent performance with accuracy ranging from 71.7% to 74.2% and AUC from 0.79 to 0.81.

We also explored the variable significance using sensitivity analysis (SA) with three measures: gradient, variance, and range. Results show that the factors that mostly affect employability are related to education and age. On the other end, the factors with least importance are region, nationality, and family structure. Apparently, the Irish labour market shows homogeneity with respect to those factors.

Finally, we built variable effect characteristics (VEC) for the three most significant variables to see further detail on how they affect the employability.

In conclusion, we find that data mining on large nationwide datasets is a technique, that helps to analyse empirically large data sources and on that basis to draw conclusion or validate conclusions, which otherwise might be arguable or intuitive.

REFERENCES

1. CSO: QNHS [Online], http://www.cso.ie/en/qnhs/
2. P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, “CRISP-DM 1.0 - Step-by-step data mining guide,” CRISP-DM Consortium, 2000
3. Menard, S. (2002). Applied Logistic Regression (2nd ed.). SAGE
4. Fisher, R., The Use Of Multiple Measurements In Taxonomic Problems. Annals of Eugenics, 1936, pp.179–188
5. McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition., 2004, Wiley Interscience

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

6. Martinez, A., Kak, A., PCA versus LDA, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2), 2001, pp.228–233
7. Tao Li, Shenghuo Zhu, and Mitsunori Ogihara. Using Discriminant Analysis for Multi-Class Classification: An Experimental Investigation. Knowledge and Information Systems, vol. 10 no.4, 2006, pp.453–72
8. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2009, http://www.R-project.org.
9. Cortez, P. "Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool". In Proceedings of the 10th Industrial Conference on Data Mining (Berlin, Germany, Jul.). Springer, 2010, LNAI 6171, 572– 583.
10. P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," Decision Support Systems, vol. 47, no. 4, 2009, pp. 547–553.
11. P. Cortez, M. Embrechts. Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences vol. 225, 2013, pp.1-17.
12. R. Kewley, M. Embrechts, C. Breneman "Data strip mining for the virtual design of pharmaceuticals with neural networks," IEEE Transactions on Neural Networks, vol. 11 (3), 2000, pp. 668–679
13. T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no.8, 2005, pp. 861–874.
14. B. Jantavan, C. Tsai, "The Application of Data Mining to Build Classification Model for Predicting Graduate Employment", International Journal of Computer Science and Information Security, vol. 11 No 10, 2013.
15. T. Mishra, D. Kumar, "Students' Employability Prediction Model through Data Mining", International Journal of Applied Engineering Research, vol. 11. No. 4, 2016, pp. 2275-2282.
16. M. Sapaat, A. Mustapha, J. Ahmad, K. Chamili, R. Muhamad, "A Classification-based Graduates Employability Model for Tracer Study by MOHE", Digital Information Processing and Communications, Springer Berlin Heidelberg, 2011, pp. 277-287.
17. J. Kirimi, C. Moturi, "Application of Data Mining Classification in Employee Performance Prediction", International Journal of Computer Applications, vol. 146,No 7, 2016, pp. 28-35.
18. Y. Alsultanny, "Labor Market Forecasting by Using Data Mining", International Conference on Computational Science, Procedia Computer Science 18, Elsevier, 2013, pp.1700-1709.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)