International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-7 Issue-1, October 2017

Measuring Factors of Employment by Classification Tree Models

A. Nachev

Abstract: ***This paper presents a case study on data mining modeling, based on classification trees. The study analyzes data from a national household survey, which provides information about Irish labour and unemployment status of the respondents. Based on trained predictive models, we address some gaps in previous studies by providing means to measure and rank the employment factors and analyze their role over the studied period. Results from experiments show that features representing age and education appear as top factors affecting the employment status. Studying further each of those by VEC analysis, we find empirically the role of their values in employment success. Measuring the model performance, we came to the conclusion, that a carefully trained classification tree can outperform neural networks trained on the same data in terms of accuracy, but underperforms neural nets in terms of AUC.***

Index Terms **: classification, data mining, labour, classification trees.**

I. INTRODUCTION

According to a study [9, 10], the impact that the economic downturn had on Ireland's labour market caused unemployment rate increase from 4.6% in 2006 to 15% in 2012. Young people were particularly hard hit with unemployment rate increasing from 9.9% per cent to 33% over that period. Authors also claim that the proportion of unemployed youths with no formal education increased over the recessionary period. The negative effect of having low levels of education, such as junior cert or less, on finding a job has become stronger since the recession; while a Post-Leaving Cert (PLC) level qualification (which includes apprenticeships) was no longer as important for unemployed youths in securing employment, most likely due to the "substantial fall" in the demand for vocationally qualified labour in construction and related sectors that took place during the recession years. Analyzing data from the Quarterly National Household Survey (QNHS) provided by the Irish Central Statistics Office's (CSO), Kelly and al. [9, 10] show that prior to the downturn, young women were more likely to be unemployed than men, a situation that has been reversed afterwards. Despite using data analysis techniques, such as non-linear decomposition models, the conclusions from that study lack of sufficient measuring of the power of the factors affecting the respondents' employment status.

This study aims to address those gaps by using decision trees as a data-mining modelling technique for building predictive models and the same QNHS data source. Along with fitting models to the data, we measure and rank the factors that affect employment status and also do variable effect characteristic analysis of those factors.

Analysis of data by the means of data mining may use variety of 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.

Most of the studies in the domain of labour and employment focus to students and graduates as target group [8, 13, 17] or to workers at organisational level for the purposes of HR management [1, 12]. Literature suggests, that the data mining modeling techniques used include decision trees, and Bayesian methods [1, 8, 12, 13, 17], ensemble methods, MLP, and SVM [13].

This paper focuses on classification trees as a modeling technique discussing issues related to building optimal trees, model performance estimation, measures factors affecting employment and analyses those factors.

The remainder of the paper is organized as follows: section II provides an overview of the data mining technique used; section III discusses the dataset used in the study, its features, and the preprocessing steps needed to prepare the data for experiments; section IV presents and discuses the experimental results; and section V gives conclusions.

II. DECISION TREES

Decision tree methods for classification and regression were developed by Breiman et al. [2]. The methodology they created was called CART (Classification And Regression Trees) and the related procedure called C4.5. Here we focus to classification trees, as the data-mining task we target is binary classification. Two ideas underpin the classification trees: recursive partitioning and tree pruning. Recursive partitioning is an iterative process of splitting the training data into partitions and then splitting it further on each of the tree branches. Dealing with cross-sectional data set, let's denote x the vector of predictor variables $x_1, x_2, ..., x_p$, and the dependent categorical variable y. The x variables can be continuous, ordinal, or binary. Objective of the first step of recursive partitioning is to split the p-dimensional space of the x variables into two non-overlapping half-spaces. In order to do so, a variable, e.g. x_i, is selected as splitter variable and a value, say s_i, is chosen as split value (or split point) for that variable. All data points of the training set with $x_i \pounds s_i$ go to one partition, and those with $x_i > s_i$ go to the other. This split forms the root node of the tree and the two branches, which lead to the next two nodes representing the split partitions (Fig. 1).

Manuscript published on 30 October 2017.
***** Correspondence Author (s)
A. Nachev*, BIS, Cairnes Business School, NUI Galway, Galway, Ireland.

Next the process continues recursively with each of the two partitions in order to be further split into smaller partitions, etc. Each split creates a new node in the decision tree. That process stops when the sub-partitions get as homogeneous or 'pure' as possible - that means data points belong as much as possible to one class only. The pure nodes are called terminal or leaf nodes. Absolutely pure might not be always possible due to the positions of the data points.

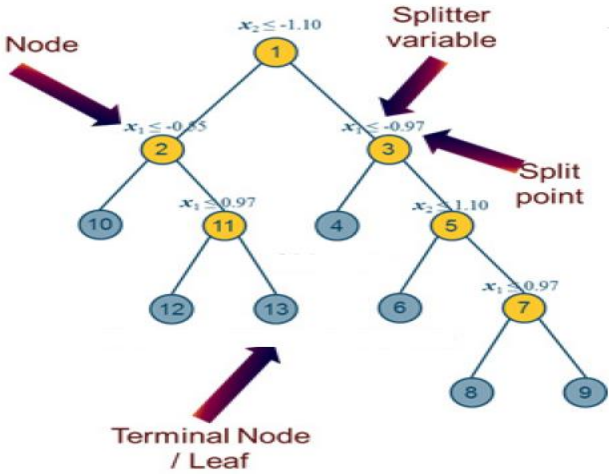

Figure 1. Example of a Decision Tree.

Two metrics of impurity of the partitions are mostly used, Gini index and information gain. Let $m=$ is the number of classes of y.

Gini index for a child node partition A that results from a split in the tree is defined by:

$$I(A) = 1 - \sum_{k=1}^{m} p_k^2 = \sum_{i \neq k} p_i p_k \qquad (1)$$

where p_k is the proportion of observations in A that belongs to class k $(k=1,2,...,m)$. For binary classifiers, Gini index takes values between *0* (all data points belong to the same class), and *0.5* (data points are equally distributed between the two classes). For multi-class, upper end is *(m-1)/m*, where all *m* classes are equally represented.

Information gain (IG) is the other most common impurity measure, based on the on the concept of entropy from the information theory. Entropy (H) is defined as

$$H(A) = -\sum_{k=1}^{m} p_k \log_2(p_k) \qquad (2)$$

Entropy for binary classification ranges between 0, where all data point belong to the same class (pure partition) and 0.5 where the two classes are equally distributed. The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain, i.e. the most homogeneous branches. The procedure first splits the dataset on a predictor variable. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the information gain, or decrease in entropy. In summary,

$$IG(parent) = H(parent) - [average\ H(children)] \qquad (3)$$

Although information gain is usually a good measure for deciding the relevance of an attribute, it has drawbacks. One of the most notables is when information gain is applied to attributes that can take on a large number of distinct values.

In order to find the best split variable at a node, the tree building algorithm starts with considering each predictor variable in turn. Every possible split is tried and considered and the best split is the one that produces highest purity of sub-partitions. The winner is choice of the root node of the tree specifying the splitter variable and split value. That process continues recursively, adding sub-nodes, thus growing the tree until further splitting cannot create purer partitions. That is a full-grown tree build by the training dataset, hence it provides best performance on classifying data from the training dataset. Trees grown by CART are binary trees in which leaf nodes are exactly one more than the number of decision nodes.

To handle categorical predictor variables, the tree building algorithm considers its values as a set and tries all possible pairs of subsets when choosing the best split. Some algorithms support binary categorical variables only and if that is the case, one have to replace those variables with several dummy ones, each of which is binary. Here we discuss binary classifiers presuming there are only two class labels, but in case of multi-class tasks, multi-way partitioning can be achieved through repeated binary splits.

Evaluating the performance of a classification tree we start by using the training set to grow the tree and then the validation set is used to assess its performance. The full-grown tree leads often to 100% accuracy in classifying the training dataset, but that tree shows lower accuracy for the validation dataset. The main reason is that that the full-grown tree overfits the training data. An overfit tree shows poor performance on new data as well. Two approaches are used to avoid overfitting, by stopping the tree to grow at certain stage or by pruning the full-grown tree back to a level where it does not overfit. A popular method to stop a tree to grow and overfit is CHAID (Chi-squared Automatic Interaction Detection), which uses chi-squared statistical test for independence to assess if splitting a node improves purity or not by statistical significance. For each decision node, the splitter predictor is the one that has the strongest association with y. If for a split the test does not show significant improvement, the split is not carried out and the tree branch is terminated. Pruning a full-grown tree is the method adopted by C4.5 and CART. C4.5 uses the training set to both grow the tree and prune it. CART uses the validation set to prune the full-grown tree by removing the 'weakest' branches that do not reduce significantly the error rate. A good approach to find the best tree is to use the complexity parameter (cp), discussed in Section IV. Regression trees for prediction tasks are composed in pretty similar manner, the major difference is that the output variable y is continuous and therefore 'purity' cannot be measured using classes. A typical measure for impurity is the sum of the squared deviations from the mean of that node, which is equivalent to the squared errors. In regression trees the value of the leaf node is determined by the average of the training data in that leaf. Decision trees are good for estimation of variable significance and variable selection in classification and regression tasks,

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

As the most important predictor variables show up at the top of the tree. In that sense decision trees don't need variable selection procedure prior to the model building as the variable subset selection is automatic and part of the splitter selection mechanism. Another advantage of decision trees is their non-linearity, which make them great performers in non-linear tasks, but that is particularly applicable in cases where discriminating between classes works well with horizontal and/or vertical splits of the data space. Decision trees are not that good in capturing diagonal or arbitrary non-linear splits of the space, where other techniques may perform better (see Fig. 2). Decision trees are also very robust to outliers.

On the other hand, a disadvantage of decision trees is their requirement to train on a large dataset in order to build a good model. They are also relatively computationally expensive, as the tree growth requires trying with multiple potential splitter variables, split values, and sorting.

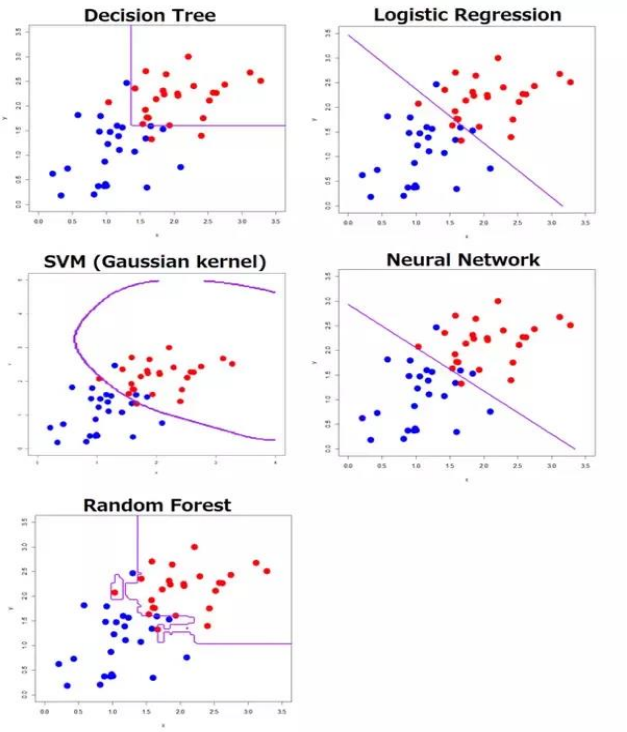

Figure 2. Example of classification techniques trained on a dataset with two attributes.

One of the reasons decision trees techniques are popular for classification and regression tasks is that they are easy understandable and interpretable and also easy to be converted to a set of classification or regression rules.

III. DATASET AND PRE-PROCESSING

This study uses the Quarterly National Household Survey (QNHS) [6] - a large-scale, nationwide survey of households in Ireland. It is designed to produce quarterly labour force estimates that include the official measure of employment and unemployment in the state. The sample data used in this study covers 2014-15, divided in four consecutive half-year terms denoted from T1 to T4. This period is selected to capture recovery form the post-2008 Irish economic downturn, in which Irish unemployment rate rose from 4.2% in 2007 to reach 14.6% in February 2012 (Fig. 3). The context of ecomonic recovery allows to explore how factors playing significant role in unemployment change in changing economic climat.

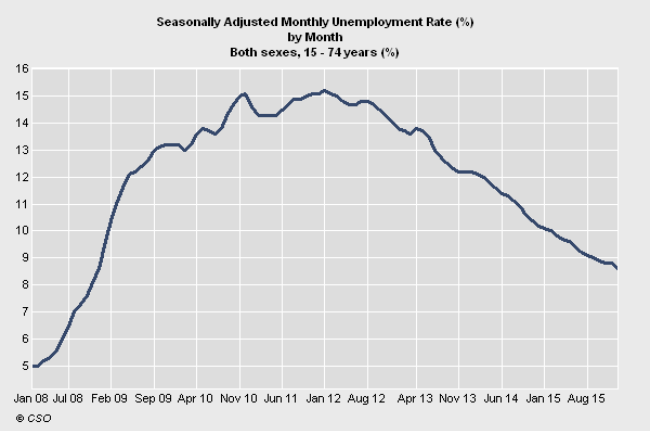

Figure 3. Irish national unemployment rate from 2008 to 2015. Source: Irish Central Statistical Office (CSO).

Originally, the four sub-datasets have sizes: T1 - 52,763; T2 - 50,515; T3 - 50,939; T4 - 45,047 observations. QNHS provides 115 variables grouped into three categories [6]:

- *Core* variables provide information about respondent's demographics, labour status, employment characteristics, atypical work, hours worked, second job, previous work experience of unemployed, search of employment, education and training, dwelling unit information.
- *Derived* variables, labour related.
- *Family unit* related.

A large number of those variables were deemed unrelated to the data-mining task and eliminated as part of the pre-processing stage. For example, variables containing information about few, but not to all respondents were discarded. Also, variables nearly identical or deemed dependent to others were eliminated due to correlation. A new binary variable ILO_BIN was added to the dataset. It was derived form the non-binary ILO and used as dependent variable for the binary classification task. After variable elimination, the original 115 were reduced to 17 in five groups, as follows:

- *Demographic*: SEX (gender); MARSTAT (marital status); NATIONAL_SUMMARY (nationality of the respondent); YEARESID_SUMMARY (years of residence in this country).
- *Education*: EDUCLEVEL (recent/ongoing education and training 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);

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

- *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)
- *Target* variable is ILO_BIN.

The dataset observations were also cleansed by discarding those with out of working age (out of 16 - 75). Also, records containing missing values were removed. Finally, the four subsets were reduced to the following sizes: T1 - 35978 records, T2 - 30409, T3 - 34240, and T4 - 28978 records.

Partitioning is a pre-processing step, which is required for building a model. Depending on the case, partitioning may break the dataset into two or three parts - training, validation, and optionally testing partitions. The model trains by fitting to the training partition presented to it, while the validation partition usually controls training by testing the model until it reaches satisfactory performance. With decision trees, the validation set is used for pruning a full-grown tree to one, which is optimal. In that way, despite the validation set is not directly involved in training, it indirectly influences the training by modifying the tree. Measuring the model performance by the validation set only is straightforward, but not data-neutral and the figures of merit obtained in that way might be quite optimistic. A separate testing partition used solely for testing and not presented during training can produce more realistic estimates of the model performance.

This study uses four dedicated test partitions for each term T1-T4, obtained by random selection of 20% of the original term dataset. The rest of observations for each term were split randomly into training and validation partitions in ratio 2:1.

IV. EXPERIMENTS AND DISCUSSION

The experiments were conducted using R environment [4, 15, 18] addressing three issues:

- Building best performing binary classification models for each of the terms T1 - T4. Performance estimators incude classification accuracy and ROC analysis. Figures of merit show performance on the test data partitions.
- Esimating variable significance for those models. Ranking predictor variables provides insight with regard to the factors affecting employment. Methodology used is a combination of sensitivity analysis and variable importance derived from the tree stricture.
- Exploring most important factors using Variable Effect Characteristic (VEC) analysis.

Models were built using the *rpat* package [16] for recursive partitioning for classification and regression trees, which implements most of the functionality of Breiman et al. CART algorithms [2]. Using the pre-processed QNHS T1-T4 training and validations partitions and tuning the tree complexity parameter (cp) we built trees with desired structure. The complexity parameter controls the tree size in the following way: If the cost of adding another variable to the decision tree from the current node is above the cp value, then tree building does not continue. In that sense, the tree construction does not continue unless it would decrease the overall lack of fit by a factor of cp. Therefore, cp is a proxy for the number of splits of the tree. It serves as a penalty term to control the tree size and is always monotonic with the number of splits. The smaller the value of cp, the more complex will be the tree (the greater the number of splits), which may cause overfitting of the model. Avoiding overfitting and selecting optimal cp involves several performance metrics, most important of which are relative error (*rerror*), cross-validation relative error (*x-val relative error*, *xerror*), and cross-validated standard deviation (*xstd*) a.k.a. standard error (SE).

The *rerror* is root mean squared error, similar to linear regression. This is the error on the observations used to estimate the model.

The *rpart* uses built-in 10-fold cross-validation (CV) procedure for building tree and *xerror* is the cross-validation relative error on the training partition.

The next section illustrates building an optimal classification tree using T1 data. Table 1 shows the region of metric values, where optimal cp is expected. Columns represent the seq. number of cp (*#*), cp value, number of splits in the tree for that cp (*nsplit*), *rerror*, *xerror*, and *xstd*. Each row represents a different height of the tree. Generally, more levels in the tree lower classification error on the training partition, but that also increases the risk of overfitting.

Choosing optimal cp that avoids overfitting is where *xerror* is minimal. That is cp = 0.514650 in row 29 of Table 1. On the other hand, Breiman et al. [2], recommend using the "1SE rule" to select cp. This is to find the minimum *xerror*, but then go up one SE because the corresponding tree is less complex. In Table 1 that is *xerror* = 0.514650 + 1*0.005568 = 0.520218, corresponding to row 16, where cp=0.000625.

The package *rpart* also provides a graphical tool for choosing the optimal cp (Fig. 4). It plots the cp of a full-grown tree on the x-axis (along with the corresponding tree size) versus the cross-validated *xerror* on the y-axis. The dashed horizontal line is drawn one standard error (1SE) above the minimum of the curve.

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

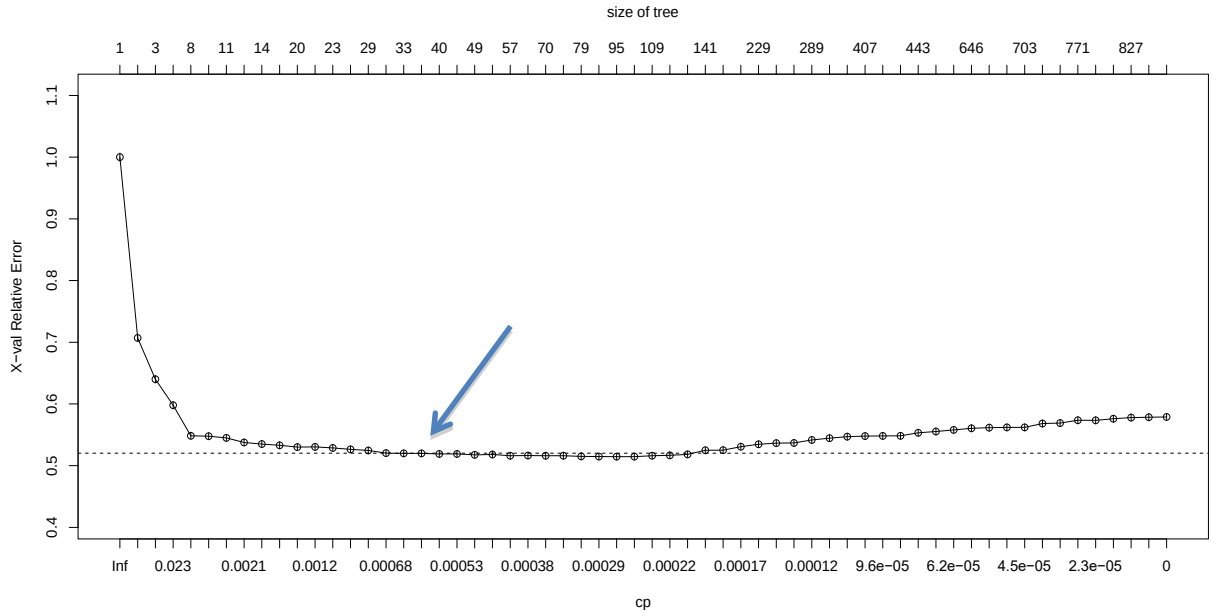

Figure 4. Plot of complexity parameter (cp) for an *rpart* fit.

The choice of cp on the plot using the "1SE rule", is where the leftmost value of the cross-validation relative error lies below the horizontal line.

Table 1. Optimal tree pruning using complexity parameter cp.

#	cp	nsplit	rerror	xerror	xstd
...					
14	7.81E-04	24	0.514	0.526447	0.005613
15	7.42E-04	28	0.511	0.524572	0.005606
16	**6.25E-04**	**31**	**0.509**	**0.520353**	**0.005590**
17	5.86E-04	32	0.508	0.519963	0.005589
18	5.66E-04	34	0.507	0.519884	0.005588
19	5.47E-04	39	0.504	0.518947	0.005585
20	5.08E-04	44	0.501	0.518947	0.005585
21	4.69E-04	48	0.499	0.517540	0.005579
22	4.30E-04	53	0.497	0.518166	0.005582
23	3.91E-04	56	0.495	0.516056	0.005574
24	3.65E-04	66	0.491	0.516369	0.005575
25	3.52E-04	69	0.490	0.516134	0.005574
26	3.44E-04	71	0.489	0.516134	0.005574
27	3.13E-04	78	0.486	0.515040	0.005570
28	2.73E-04	87	0.484	0.514728	0.005569
29	**2.66E-04**	**94**	**0.482**	**0.514650**	**0.005568**
30	2.60E-04	99	0.480	0.514650	0.005568
31	2.34E-04	108	0.478	0.516056	0.005574
14	7.81E-04	24	0.514	0.526447	0.005613

...

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 typically underrepresented, such as the class representing unemployed respondents. For such applications, sensitivity and specificity can be more relevant performance estimators. In order to address the accuracy deficiencies, we did Receiver Operating Characteristics (ROC) analysis [7]. 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. Figure 5 shows ROC curve of a model built on T1 data with cp=0.000625.

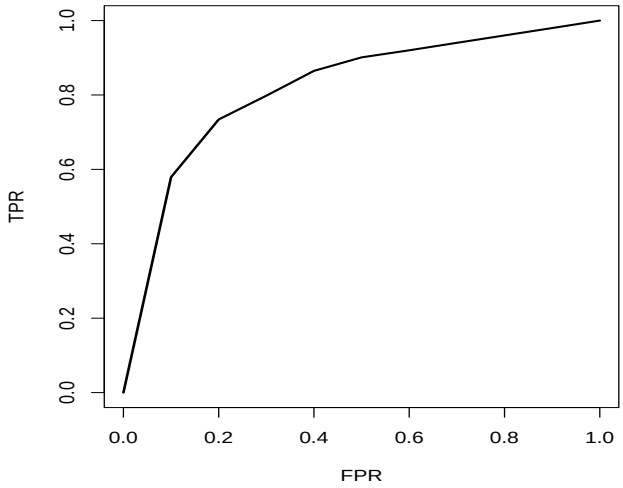

Figure 6. ROC curve of classification tree trained on T1 QNHS data and complexity cp=0.000625.

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

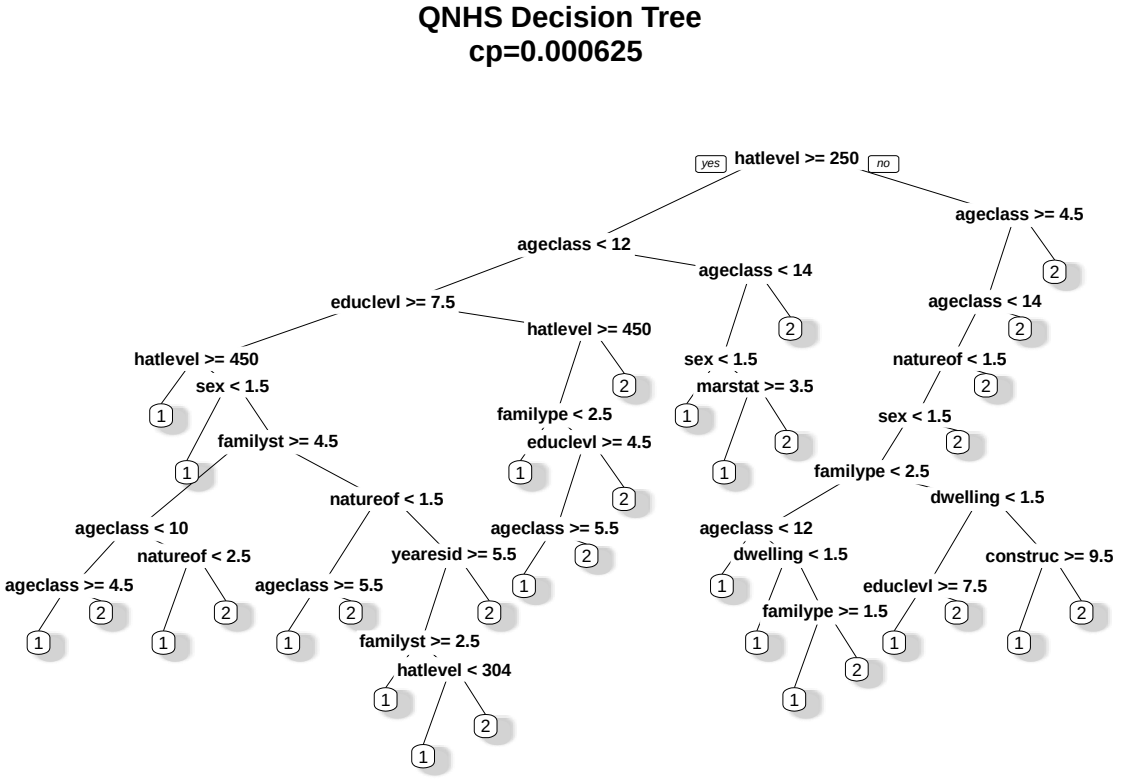

Figure 6. Classification tree trained on T1 QNHS data and complexity cp=0.000625. Classes represent employment status 1-yes, 2-n

Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test 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 test. The area under the ROC curve (AUC) is a common model performance metric. AUC represents classifier performance over all possible threshold values, i.e. it is threshold independent.

A decision tree trained on T1 data with complexity parameter cp=0.000266 provides ACC=76.932% and AUC= 0.819, which shows the model more accurate than one based on MLP neural networks and trained on the same data, but with regard to the metric AUC, neural networks provide better results (NN_ACC=76.49%, NN_AUC=0.842) [14].

Tree with complexity parameter cp=0.000625 is illustrated in Figure 5. It provides ACC=76.431% and AUC=0.813, which is negligibly worse in performance than the model mentioned above, but at the same time is much simpler and easier for interpretation. Even a simpler tree with cp=0.00164 was created in order to identify most important predictor variables, which appear as splitter variables on the nodes (Fig. 7). Generally, with decision trees the closer a variable to the root, the more important it is for the model. According to Fig. 7, most important variables are HATLEVEL and EDUCLEVEL, both related to education and training, and AGECLASS representing age.

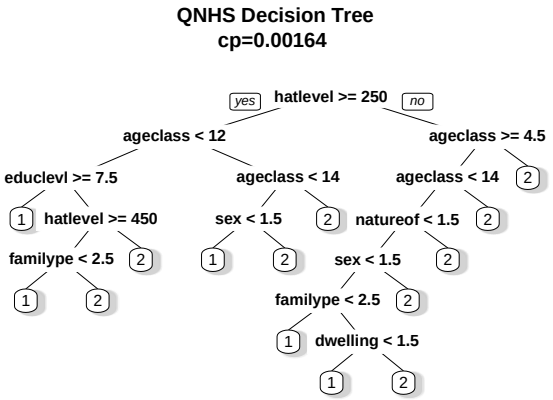

Figure 7. Classification tree trained on T1 QNHS data with complexity cp=0.00164. Classes represent employment status 1-yes, 2-no.

Next in importance is the variable SEX, indicating that gender also influences employment status. As mentioned by Kelly et al. [9, 10], prior to the economic downturn, young women were more likely to be unemployed than men, but then the situation has been reversed. One explanation in that regard might be demand for workers in temporary expanding sectors, like construction industry, where men were more likely to be employed. Apart from analysing the splitter variables for the purpose of identifyng their importance, the *rpart* package offers summary information about the model,

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

part of which is a list of variables with weights representing their importance, the sum of which is 100 (Table 2). It is evident again, that education and training variables along with age are the primary factors influencing the respondent's employment status.

Table 2. Weighted variable importance for optimal tree, cp=0.000625.

#	*Variable*	*Weight*
1	AGECLASS	25
2	HATLEVEL	25
3	HATFIELD	23
4	EDUCLEVEL	15
5	FAMILYPERSON_SUMMARY	2
6	CONSTRUCTIONDATE	1
7	FAMILYSTRUCTURE_SUMMAR	1
8	NATUREOFOCCUPANCY	2
9	SEX	3
10	NUMBEROFROOMS	0
11	MARSTAT	2
12	FAMILYTYPE_SUMMARY	0
13	DWELLINGUNIT	1
14	YEARESID_SUMMARY	0
15	NATIONAL_SUMMARY	0
16	REGION	0

In order to rank variable significance, we also used the sensitivity analysis (SA) method proposed by Kewley et al. [11], which varies each input variable x_a through its range with L levels from the minimum to the maximum value. Given x_{a_j} denotes the j-th level of input x_a and $\hat{y}$ denotes the value predicted, significance can be measured by variance measure [3, 4, 10]:

$$S_v = \sum_{j=2}^{L} (\hat{y}_{a_j} - \bar{y}_{a_{j-1}})^2 / (L-1) \qquad (3)$$

where $\bar{y}_a$ denotes the mean of the responses [3, 5, 13]. This method was initially proposed for neural networks, but it is applicable to any supervised learning technique, in particular classification trees. Figures 8 a) - d) show ranking of variable significance in each of the four periods T1 - T4. It is evident, that top three most significant variables are once again related to the age and education. Minor variances over the periods change their place in the rank, but they are always the top three. Being classification technique neutral, the SA ranking of variable importance solidifies the conclusions from the previous approaches. Looking at the relative weight of the top three factors, we can arrive to the conclusion that their collective influence to the employment status is about two thirds of all.

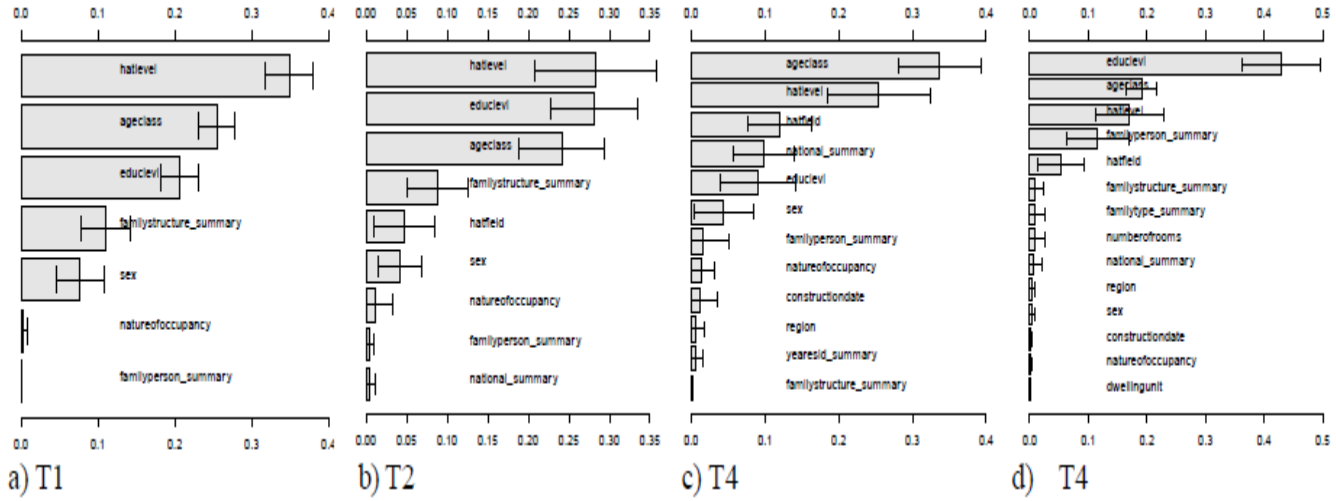

Figure 8. Variable significance for optimal tree, cp=0.000625 using variance sensitivity measure. Each figure represents a term from T1 to T4.

Further to the sensitivity analysis, Variable Effect Characteristic (VEC) analysis shows how models learn from data, and how the predictor variable values contribute to discrimination between the classes.

Following notation from (3), within the range of values of an input x_a with L levels from the minimum to the maximum, we can plot x_{a_j} on the x-axis versus responses $\hat{y}_{a_j}$ on the y-axis. Between two consecutive x_{a_j} values, the VEC plot uses a line as interpolation for continuous values and a horizontal segment for categorical data.

Figure 9 shows VEC curves for AGECLASS, HATLEVEL, and EDUCLEVEL for T1 to T4 data with L=6. Since the experiments ran 10 times per VEC curve, vertical results were averaged and whiskers on the plots show the confidence intervals. Analyzing VEC curves, we can see the role of each predictor. First, all plots suggest that the relations between variable values and their contribution to the employment status are non-linear. For instance Fig. 9 a) shows that the AGECLASS curve is convex with peak at values 6-8 corresponding to age 25-39 (young to mid-age), which together with good education and training appear as best combination for employment, particularly in the context of high unemployment rate in the T1 period.

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

Such a hypothesis is easy to come about by common sense and practical experience, but a result derived experimentally from a model trained on empirical data makes it solid evidence justifying the hypothesis. Looking at Fig. 9 b)-d), it is visible how the peak of employed age flatten comprising a wider age band with values 6-12 (young to mature), which makes the group of mature employees with experience more competitive in the context of a recovering economy and lower unemployment rate. Overall, the age class 4 (young), usually inexperienced youngsters, is not well presented in the 'employed' class. Similarly, the age group just before retirement is not well employed either.

The two most important variables related to education and training, EDUCLEVEL and HATLEVEL, represent recent or ongoing education/training (within the last 4 weeks of the interview) and completed, respectively. All plots are ascendant with few variances in scale and shape. For example, from plots e)-h) it is evident that in respondents with completed primary and lower secondary education/training (HATLEVEL < 200) are very poorly employed, in contrast of those with completed third level education and equivalent training (Bachelor degree and above, HATLEVEL > 600). It is also visible that secondary education plays moderate role in employment forming a separate group of respondents. Similar characteristics show the EDUCLEVEL plots, but in contrast to HATLEVEL, that employment factor forms two groups of respondents - those with high training (third level education and equivalent training) and all the rest - no difference between low and moderate education / training. Comparing the periods T1-T4, one can arrive to the conclusion that in bad economic climate with high unemployment rate the role of education and training is greater that in a period of well recovered economy, where those factors soften. More VEC curves can be plotted and interpreted in order to show the role of other factors, but the three most important discussed above illustrate the potential of that technique.

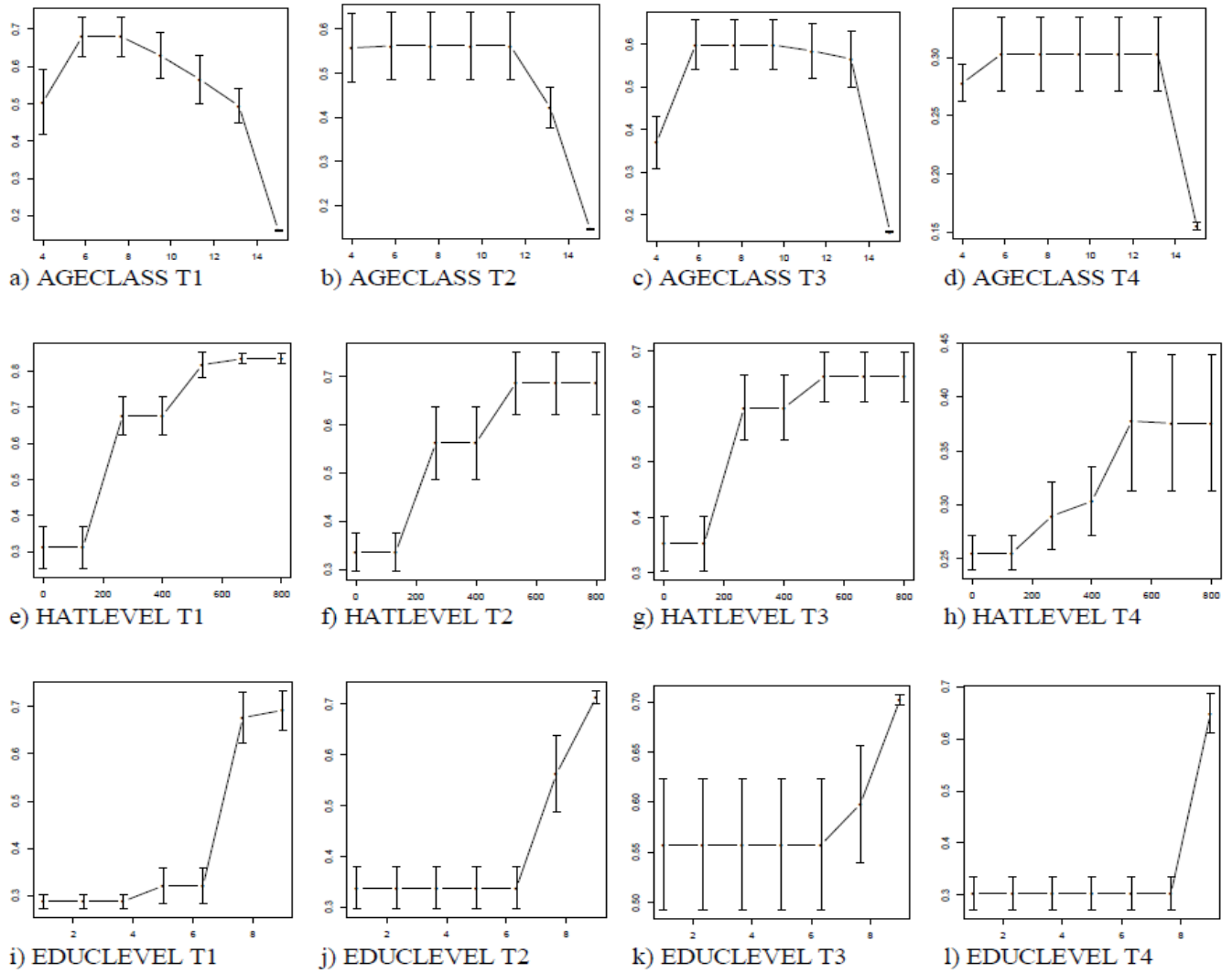

Figure 9. Vertically averaged VEC curves (points and whiskers) for AGECLASS, HATLEVEL, and EDUCLEVEL for terms T1-T4.

V. CONCLUSION

This paper presents a case study that uses classification trees to analyse empirically employment factors and their role in the Irish labour market for a certain period. It addresses some gaps in previous research, focusing to measuring factors and providing inshight with regard to their role in employment.

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

The study uses data from the Quarterly National Household Survey (QNHS), provided by the Irish Central Statistical Office. As part of the data pre-processing, the partitioning procedure used dedicated testing sets, which are both neutral to the models and provide reliable performance metrics.

Series of experiments were conducted in order to build models avoiding overfitting. Results show that this can be achieved at tree complexity parameter cp=0.000625. Measuring the model performance we found that the classification tree models can outperform those based on neural networks in terms of prediction accuracy, but underperform in terms of AUC.

Analyzing variable importance for the model, we estimated the factors that affect the employments status of respondents. By ranking those factors using two techniques, we explored the most important ones, all related to age and education and training. Further detail was given by VEC analysis, which reveals how values of those factors contribute to the employments status.

In conclusion, we find, that classification trees are well performing data mining technique with potential to derive knowledge form labour data and to validate empirically hypotheses in the area.

REFERENCES

1. Alsultanny, Y. Labor Market Forecasting by Using Data Mining, International Conference on Computational Science, Procedia Computer Science 18, Elsevier, 2013, pp.1700-1709.
2. Breiman, L., Friedman, J., Olshen, R.A., Stone, C. Classification and Regression Trees, Wadsworth, Belmont, CA, 1984
3. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J. Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems, vol. 47, no. 4, pp. 547–553, 2009.
4. 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, LNAI 6171, 572– 583, 2010.
5. Cortez, P., Embrechts, M. Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences vol. 225, 2013, pp.1-17.
6. CSO: QNHS [Online], http://www.cso.ie/en/qnhs/
7. Fawcett, T., An introduction to ROC analysis, Pattern Recognition Letters 27, No.8, 861–874, 2005
8. Jantavan, B., Tsai, C. 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.
9. Kelly, E., McGuinness, S. (2014, online), Impact of the Great Recession on Unemployed and NEET Individuals’ Labour Market Transitions in Ireland, Economic Systems. http://dx.doi.org/10.1016/j.ecosys.2014.06.004
10. Kelly, E., McGuinness, S., O’Connell, P., Haugh, D., Pandiella, A. (2014), Transitions In and Out of Unemployment among Young People in the Irish Recession, Comparative Economic Studies, 56: 616-634.
11. Kewley, R., Embrechts, M., Breneman, C. Data strip mining for the virtual design of pharmaceuticals with neural networks. IEEE Transactions on Neural Networks, vol. 11 (3), pp. 668–679, 2000
12. Kirimi, J., Moturi, C. "Application of Data Mining Classification in Employee Performance Prediction", International Journal of Computer Applications, vol. 146,No 7, 2016, pp. 28-35.
13. Mishra, T., D. Kumar, "Students' Employability Prediction Model through Data Mining", International Journal of Applied Engineering Research, vol. 11. No. 4, 2016, pp. 2275-2282.
14. Nachev, A., (2017) 'Using Multi-Layer Perceprons for Analysis of Labour Data', In Proc. of International Conference Artificial Intelligence, ICAI’17, Las Vegas, 17-20 Jul, pp.223-229.
15. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org, 2009.
16. https://cran.r-project.org/web/packages/rpart/rpart.pdf
17. Sapaat, M., 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.
18. Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., ROCR: visualizing classifier performance in R., Bioinformatics 21(20):3940-3941, 2005.

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