International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 18 (2018) pp. 13525-13535
© Research India Publications. http://www.ripublication.com

Analysis of Employment Data Using Support Vector Machines

Anatoli Nachev

BIS, Cairnes School of Business and Economics,
National University of Ireland, Galway, Ireland.

Teodosi Teodosiev

Department Computer Science, Shumen University, Shumen, Bulgaria.

Abstract

This paper explores predictive abilities of support vector machines used to explore factors affecting employment rates. The study analyses data from a national household survey, which provides information about Irish labour and unemployment status of respondents over a period. Based on trained predictive models, we address some gaps in previous studies by providing means to identify, measure and rank the employment factors and analyse their role over the studied period. Using predictive models the study explores interactions between variables and provides insight on their role in employment. We estimated variable significance, reduced data dimensionality, and selected optimal model parameters for the support vector machines. Predictive abilities of those models were estimated by accuracy, ROC analysis, and AUC. In order to provide unbiased results, we applied rigorous testing procedure combining two testing partitions, cross-validation and iterations. Our results show that support vector machines outperform other modelling applied to the same data. We also did VEC analysis of the most significant employment factors and explored their characteristics. The methodology for processing labour data proposed by this study and the conclusions made provide a way to derive empirical knowledge applicable to the specific data we use.

Keyword: Data mining, classification, support vector machines, labour market.

INTRODUCTION

In recent years, there has been growing interest in active management of the labour market and development of policies and measures directed towards unemployment. Studying that area goes to areas, such as providing efficient job seeking services and counselling, upgrading and adapting skills programmes, job creation programmes, etc.

With growth of available social information in large datasets and availability of powerful computer-based processing tools and techniques for business analytics and data mining, it becomes important to derive empirical knowledge from those data sources in order to contribute to theoretical body of knowledge in the field. Studies in the area provide value not only by disclosing facts and relationships between factors and variables, but also provide a useful empirically-based validation tool for theoretical considerations in that domain.

According to Kelly and al. [18], [19], 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. [18], [19] 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 analyses data from the QNHS dataset, gathered by the Irish national survey, conducted regularly, in order to identify the role of demographic characteristics, education, dwelling unit information, and family status of the Irish population in the level of employability. We also drill down in each of these factors, providing further analysis of their role and structure. The analysis uses modelling based on support vector machines, which has been trained and tested using the survey data. 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 an effective supervised learning method for building predictive models.

A number of works in the domain of labour and employment have been done recently. Majority of those studies focus to students and graduates as target group [20], [21], [22] or to workers at organisational level for the purposes of HR management [23], [24]. Literature suggests, that the data mining modeling techniques used include decision trees, and Bayesian methods [20], [21], [22], [23], [24], ensemble methods, MLP, and SVM [21].

This paper focuses on support vector machines as a modeling technique discussing issues related to building models, such as choosing optimal parameters, performance estimation, measuring factors affecting employment and analysing of those factors.

The remainder of the paper is organized as follows: section II provides an overview of the support vector machines as a data mining technique; 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 discusses experimental results; and section V gives conclusions.

SUPPORT VECTOR MACHINES

Support vector machines have grown in status over the past decade due to the satisfactory results returned over a diverse range of fields. SVM are data analysis techniques categorised within the domain of supervised machine learning [1], [2], whereby the learning process results in a function being contingent on the supervised training data. Through this supervised machine learning process, the algorithm returns either a classification function, or a regression function. A support vector regression procedure suggests an optimal tradeoff between complexity and learning ability in order to achieve a strong generalization of accuracy [25].

For a two-class, separable training data set, such as the one in Figure 1, there are lots of possible linear separators. Intuitively, a decision boundary drawn in the middle of the void between data items of the two classes seems better than one which approaches very close to examples of one or both classes. While some learning methods such as the perceptron algorithm find just any linear separator, others, like Naive Bayes, search for the best linear separator according to some criterion. The SVM in particular defines the criterion to be looking for a decision surface that is maximally far away from any data point. This distance from the decision surface to the closest data point determines the margin of the classifier. This method of construction necessarily means that the decision function for an SVM is fully specified by a subset of the data points, which defines the position of the separator. These points are referred to as the support vectors. Figure 2 shows the margin and support vectors for a sample problem. Other data points play no part in determining the decision surface that is chosen.

SVM can be formalized as follows. Training data D of n samples is a set of pairs of data points $\vec{x}_i$ (p-dimensional vectors) and class labels y_i where -1 indicates one class; +1 the other class.

$$D = \{(\vec{x}_i, y_i) | \vec{x}_i \in \Re^p, y_i \in \{-1, +1\}\}_{i=1}^{n} \quad (1)$$

During training a SVM builds a decision boundary that separates the classes. The decision boundary is a p-1 – dimensional hyperplane (a line in the 2D case, a plane in the 3D case, etc.). A decision hyperplane can be defined by a normal vector $\vec{w}$ perpendicular to the hyperplane and a term b. The vector $\vec{w}$ is often called weight vector. The term b specifies the choice of hyperplane among all perpendicular to the normal vector. Because the hyperplane is perpendicular to the normal vector, all points $\vec{x}$ on the hyperplane satisfy

$$\vec{w}^T \vec{x} + b = 0 \quad (2)$$

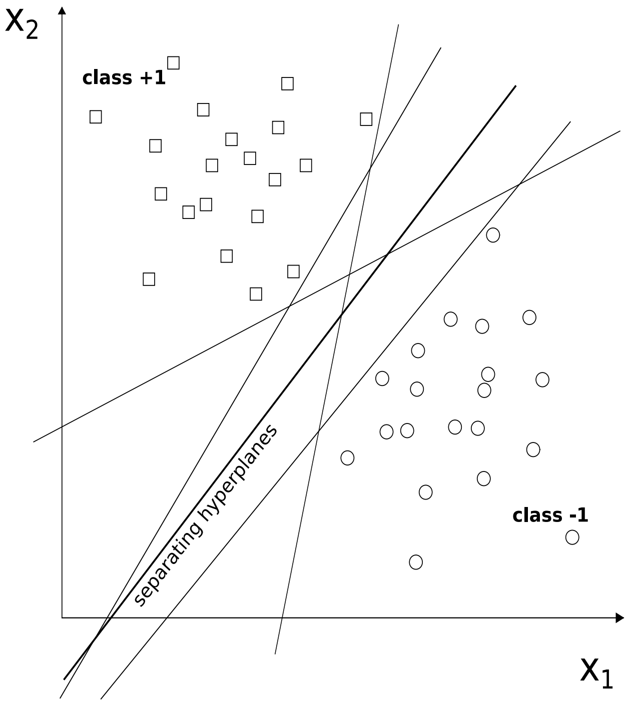


Figure 1: Separating lines for a two-class separable dataset.

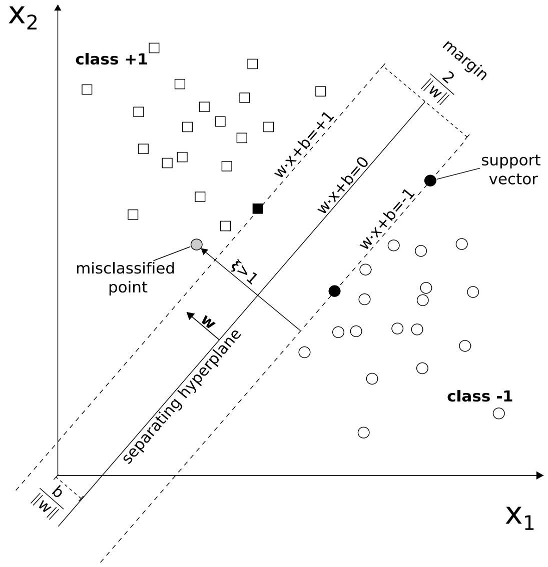


Figure 2: Geometry of support vector machines.

Data points would fall into one or another side of the decision hyperplane turning the above equality into inequality, therefore the decision function of a linear SVM classifier can be defined as

$$f(\vec{x}) = sign(\vec{w}^T \vec{x} + b) \quad (3)$$

Class labels are +1, -1. The points closest to the separating hyperplane are called support vectors. The margin of a classifier is the maximum width of the band that can be drawn separating the support vectors of the two classes. It can be shown that maximizing the margin is the following minimization problem: find $\vec{w}$ and b, such that

$\frac{1}{2}\vec{w}^T\vec{w}$ is minimized and for all $\{(\vec{x}_i, y_i)\}$ and

$$y_i(\vec{w}^T\vec{x} + b) \geq 1 \tag{4}$$

This task is optimization of a quadratic function subject to linear constraints. The solution of that problem involves constructing a dual form of the optimization problem where a Lagrange multiplier α_i is associated with each constraint $y_i(\vec{w}^T\vec{x} + b) \geq 1$ in the primal problem. The dual problem is: find $\alpha_i, \ldots, \alpha_N$, such that

$$\sum_{\mathrm{i}} \alpha_i - \frac{1}{2}\sum_{\mathrm{i}}\sum_{\mathrm{j}} \alpha_i\alpha_j y_i y_j \vec{x}_i^T\vec{x}_j \tag{5}$$

is maximized and $\sum_{\mathrm{i}} \alpha_i y_i = 0;\ \alpha_i \geq 0$ for all $1 \leq i \leq N$.

A solution of that problem allows building the decision hyperplane:

$$\vec{w} = \sum_i \alpha_i y_i \vec{x}_i\ ; \quad b = y_k - \vec{w}^T\vec{x}_k \tag{6}$$

for any $\vec{x}_k$, such that $\alpha_k \neq 0$.

Most Lagrange multipliers found by the optimization problem are zero. Each non-zero indicates that it corresponds to a support vector. The classification function (2) can be presented in the form

$$f(\vec{x}) = sign(\sum_i \alpha_i y_i \vec{x}_i^T\vec{x} + b) \tag{7}$$

The above formulas that contain vectors also use dot product operation between them.

The simplest way to divide two classes is with a straight line in 2D, flat plane in 3D or an (N-1)–dimensional hyperplane in an N-dimensional attribute space. Sometimes, however, such a separation is impossible (Fig 3).

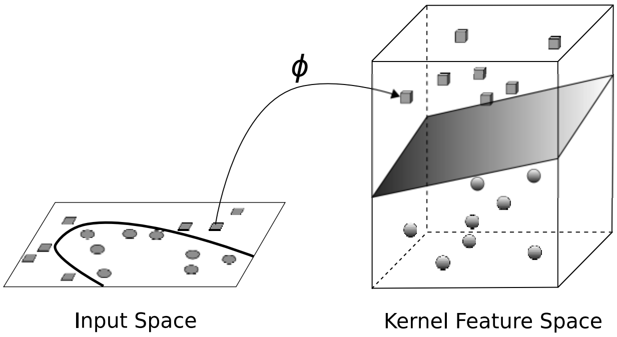


Figure 3: The kernel trick: a linearly inseparable input space can be mapped to a higher dimensional space, which is linearly separable.

Instead of fitting nonlinear curves (hyper-surfaces) to the data, an SVM can handle this using a kernel function that maps the data to a different higher dimensional space where a hyperplane can be used to do the separation. Indeed, if there are two data attributes (2D data points) and data set is not linearly separable by a line, the kernel function can add a third attribute in order to map the points into 3D, so that the data set could be linearly separable by a flat plane in 3D. It can be generalised that the kernel function transforms the data into a higher dimensional space to make separation by hyperplanes possible.

The kernel function can be defined as

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i)\Phi(\vec{x}_j) \tag{8}$$

where $\Phi(\vec{x})$ maps the vector $\vec{x}$ to some other Euclidean space. The dot product $\vec{x}_i \cdot \vec{x}_j$ in the formulas above is replaced by $K(\vec{x}_i, \vec{x}_j)$ so that the SVM optimization problem in its dual form can be redefined as: maximize (in α_i)

$$\begin{gathered}\tilde{L}(\alpha) = \sum_{\mathrm{i}} \alpha_i - \frac{1}{2}\sum_{\mathrm{i}}\sum_{\mathrm{j}} \alpha_i\alpha_j y_i y_j K(\vec{x}_i, \vec{x}_j)\,, \text{s.t.} \\ \sum_{\mathrm{i}} \alpha_i y_i = 0\,,\ \alpha_i \geq 0, \quad 1 \leq i \leq N\end{gathered} \tag{9}$$

Various kernel functions can be used with SVM and perhaps their number is infinite. But a few of them have been found to work well for a wide variety of applications. These are:

- Linear: $K(\vec{x}_i, \vec{x}_j) = \vec{x}_i^T\vec{x}_j$ (10)
- Polynomial: $K(\vec{x}_i, \vec{x}_j) = (\gamma\vec{x}_i^T\vec{x}_j + r)^d$, $\gamma > 0, r > 0$ (11)
- Gaussian RBF:

$$\begin{gathered}K(\vec{x}_i, \vec{x}_j) = exp(\sigma\|\vec{x}_i - \vec{x}_j\|^2) \\ \sigma > 0\end{gathered} \tag{12}$$

- Sigmoid (hyperbolic tangent):

$$\begin{gathered}K(\vec{x}_i, \vec{x}_j) = tanh(\gamma\vec{x}_i^T\vec{x}_j + r) \\ \gamma > 0, r > 0\end{gathered} \tag{13}$$

- Laplace: $K(\vec{x}_i, \vec{x}_j) = exp(-\frac{\|\vec{x}_i - \vec{x}_j\|}{\sigma})$ (14), $\sigma > 0$

Ideally, an SVM analysis should produce a hyperplane that completely separates the feature vectors into two non-overlapping groups. However, perfect separation may not be possible, or it may result in a model in so high dimensional space that the model does not generalize well. To allow some flexibility in separating the classes, the soft-margin SVM proposed by Cortes and Vapnik [3] permit some misclassifications. The method chooses a hyperplane that splits data points as clean as possible while still maximizing the distance to the nearest cleanly split points. The method introduces slack variables ξ_i in $y_i(\vec{w}^T\vec{x}_i + b) \geq 1 - \xi_i$, $1 \leq i \leq n$, which measure the degree of misclassification of the points $\vec{x}_i$. If a training example lies on the 'wrong' side of the hyperplane, the corresponding $\vec{w}^T\vec{x} + b$ is greater than 1. Therefore, the primal form of the optimization problem is:

$$\begin{gathered}\min_{w,\xi,b}\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\}, \qquad \text{s.t.} \\ \forall_{i=1}^{n}: y_i(\vec{w}^T\vec{x}_i + b) \geq 1 - \xi_i;\ \forall_{i=1}^{n}: \xi_i > 0\end{gathered} \tag{15}$$

The factor C in the formula is a parameter that represents the cost of misclassification. A small value of C will increase the number of training errors, while a large C will lead to a behavior similar to that of a hard-margin SVM. In that sense the cost parameter C that controls the trade-off between allowing training errors and forcing rigid margins.

The soft-margin optimization problem along with the constraint can be solved using Lagrange multipliers (as before) so that in a dual form it can be formulated as follows: minimize

$$\tilde{L}(\alpha) = -\sum_{i=1}^{n}\alpha_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j)\text{, s.t.}$$
$$\forall_{i=1}^{n}: \sum_{i=1}^{n}\alpha_i y_i = 0,\ \forall_{i=1}^{n}: 0 \le \alpha_i \le C$$

The advantage of the dual form is that the slack variables vanish, with the parameter C appearing only as an additional constraint on the Lagrange multipliers.

The SVM method can also be applied to the case of regression. A version of SVM for regression, called support vector regression (SVR), was proposed by Drucker et al. [4]. The basic idea of SVR is that a non-linear function learns by a linear learning method in a kernel-induced higher dimensional space. Similarly to how SVM classification ignores data points that are not support vectors, the SVR depend on a small subset of training data points.

The SVM's major advantage lies with their ability to map variables onto an extremely high feature space. This, in essence facilitates a means for the exploration of nonlinear kernel-based classifiers [5], [6], however they have been discovered to not favour large datasets, due to the demands it imposes on virtual memory, and the training complexity resultant from the use of such a scaled collection of data [7], [8].

Work from Fei et al. [9] highlighted three "crucial problems" in the use of support vector machines. These are attaining the optimal input subset, correct kernel function, and the optimal parameters of the selected kernel, all of which are prime considerations within this study. Multiple authors also echoed sentiments of kernel selection problems [10], [11], [12], which further indicated the importance of this factor for this research.

DATA SET AND PRE-PROCESSING

In order to see SVM potential to explore the role of different factors that affect employment in nation-wide scale, we selected data from the Quarterly National Household Survey (QNHS) [13] over the period Jan 2014 to Dec 2015, characterized with a clear trend of economic recovery from the post-2008 Irish economic downturn and also showing dropping unemployment rates. The QNHS data was divided in four consecutive half-year terms denoted as T1 to T4 and having the following size: T1 - 52,763 records; T2 - 50,515; T3 - 50,939; and T4 - 45,047 records.

Originally, the dataset contains 115 variables divided into 3 categories: 104 core variables; 7 derived variables applicable to labour market analysis; and 4 derived variables for family unit analysis. As part of the data pre-processing stage, some of these variables were eliminated, because they were seen as not relevant to the data mining goal and the modelling itself. The variables were reduced to 17, grouped as demographic, educational, representing dwelling unit information, and related to family status. A summary, containing variable names, brief meaning, and sequential number is listed below.

Demographic: SEX (gender, #1); MARSTAT (marital status, #2); NATIONAL_SUMMARY (nationality of the respondent, #3); YEARESID_SUMMARY (years of residence in this country, #4).

Education: EDUCLEVL (education level, #5); HATLEVEL (highest level of education successfully completed, #6) HATFIELD (field of highest level of education successfully completed, #7);

Dwelling unit information: DWELLINGUNIT (type of dwelling the respondent lives in, #8); NUMBEROFROOMS (number of rooms, #9); CONSTRUCTIONDATE (construction date of the dwelling, #10); NATUREOFOCCUPANCY (nature of occupancy of the dwelling, #11);

Technical items related to interview: REGION (region of household, #12); AGECLASS (age class of the respondent, #13);

Family status: FAMILYTYPE_SUMMARY (type of family, #14); FAMILYPERSON_SUMMARY (person role within the family, #15); FAMILYSTRUCTURE_SUMMARY (summary of family type, 16)

Finally, the target variable is ILO_BIN, which represents the respondent's employment status.

Further to the initial variable reduction, the dataset underwent cleansing, which eliminated records not relevant to the data mining task, for example respondents of age below 16 or above 75. Records containing missing values were removed as well.

The data was then broken into four subsets, each representing a period from T1 to T4, with the following size: T1 - 35978; T2 - 30409; T3 - 34240; T4 - 28978 records, respectively.

Each of the subsets was then divided into training, validation, and testing partitions. The testing partition contains 20% randomly selected records and the rest of 80% was used for training and validation partitioned in ratio 2:1, split automatically and randomly during the training process.

We also did correlation analysis of the variables, the results of which was used for variable selection in the modelling phase discussed in the following section.

RESULTS AND DISCUSSIONS

In order to process the QNHS dataset using data mining techniques requires building as accurate models as possible, which fit to the data very well. Using these models allows for exploring the variables involved in building their knowledge. We used R environment [14], [15], [16] to create SVM models, to tune their parameters, to analyse variable importance, and build VEC curves.

Modelling and performance estimation:

Building a robust SVM model requires a methodology, which avoid pitfalls, such as poorly validated results, which can be even misleading in some cases. One issue that should be addressed is applying double-testing approach. The modelling techniques often employ training and validation partitions without using additional test partition. The validation partition does test during the model training, thus driving the training process towards minimising the error. Having figures of merit based on the validation results only, however, might be risky for some modelling techniques, as the model hyper-parameters adapt slowly to the specifics of the validation partition and when new unseen before data is presented to the model, usually the prediction accuracy is lower than suggested by the validation. Involving additional testing partition selected from the original data and not presenting it to the model during the training allows for a final neutral estimation of the model performance, which is much more realistic. In our experiments we applied double testing by using both validation and test partitions.

Secondly, the SVM models can be non-deterministic, which means that the randomness in partitioning for each run may result in variance of results. Thus, some runs may provide too optimistic results in some cases or too pessimistic in others. To address this issue, applied 3-fold cross-validation (CV) for each run and repeated each run 10 times recording average performance.

Thirdly, as the SVM are binary classifiers, the primary figure of merit is their accuracy of prediction (ACC). For a given operating point of a classifier, the ACC is the total number of correctly classified instances divided by the total number of all available instances. ACC, however, may not be reliable estimator if one or more classes are underrepresented in the dataset. In cases of such unbalancing, sensitivity and specificity can be more relevant performance estimators. In order to address potential ACC deficiencies, we did Receiver Operating Characteristics (ROC) analysis [17] for each SVM model. 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 performance metric. AUC represents classifier performance over all possible threshold values, i.e. it is threshold independent.

Building SVM fit on T1 with the default modelling settings: Gaussian RBF kernel, C=3, sigma= 0.0513, we find average $ACC_{SVM} = 76.9\%$ and $AUC_{SVM} = 0.830$, which are close to those of neural networks $ACC_{NN} = 76.7\%$ and $AUC_{NN} = 0.842$ [27]. With default parameters, SVM slightly outperforms NN in accuracy, but underperforms in AUC.

Figure 4 illustrates ROC curves of 10 SVM models on T1 data with default parameters. It shows overlapping all curves into one with no variance, which indicates consistency in training.

SVM kernels and parameters:

Performance of an SVM model largely depends on choice of kernel, tuning its parameters, and selecting value for the cost parameter C. Finding the optimal kernel and parameter values is an empirical task, as they are dataset specific and strongly depend on the nature of the task solved. No theory can suggest what would be the optimal choices, although some recommendations exist.

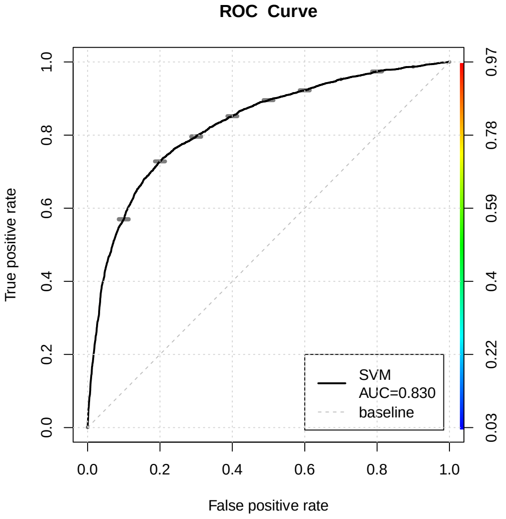
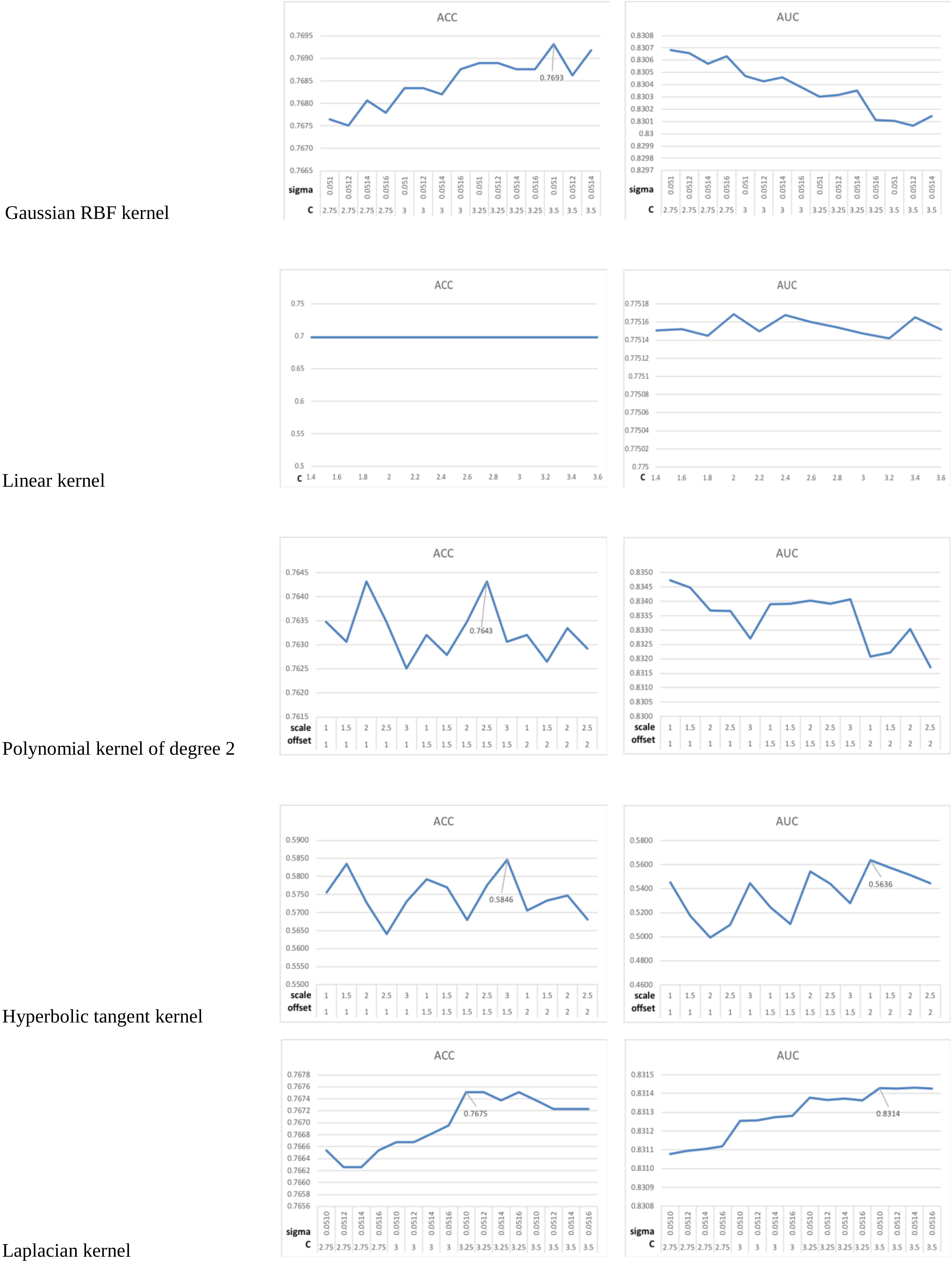


Figure 4: ROC curves of 10 SVM models. Black line represents average performance. Standard deviation bars measure variance.

The cost parameter C is a penalty factor that can control the tradeoff between the training error and model complexity, which is the number of support vectors. If C is too large, we have high penalty for non-separable data points and many support vectors, which in fact which turns a soft-margin SVM into hard-margin SVM. This leads to overfitting. On the other extreme when C is zero, we have no penalty for misclassifications, few support vectors, and model underfitting. In order to cast a broad catchment area in search of the best performing SVM, five kernels were tested: linear; both polynomial with degree 2 and parameter intervals scale=[0,5], and offset=[0,5]; sigmoid (hyperbolic tangent) with scale=[0,5], and offset=[0,5]; both Gaussian RBF with sigma=[0,5]; and Laplace with sigma=[0,5]. Also, the C parameter was explored in the interval [0,10].

We applied pattern search technique, a.k.a. compass search or a line search. It starts at the center of the search range and makes trial steps in each direction for each parameter. If the model performance improves, the search center moves to the new point and the process is repeated. If no improvement is found, the step size is reduced, and the search is tried again. Figure 5 illustrates SVM performance with five kernels and parameter values close to their optimum. Results show that best accuracy SVM uses Gaussian RBF kernel with sigma=0.051 and C=3.5. Next in performance are Laplacian kernel, polynomial, linear, and sigmoid. For the rest of our experiments we used Gaussian RBF kernel with the optimal parameters values mentioned above.



Gaussian RBF kernel

Linear kernel

Polynomial kernel of degree 2

Hyperbolic tangent kernel

Laplacian kernel

Figure 5: SVM performance using Gaussian RBF, linear, polynomial, hyperbolic tangent, and Laplacian kernel varying their parameters. Figures of merit are accuracy (ACC) and area under the curve (AUC).

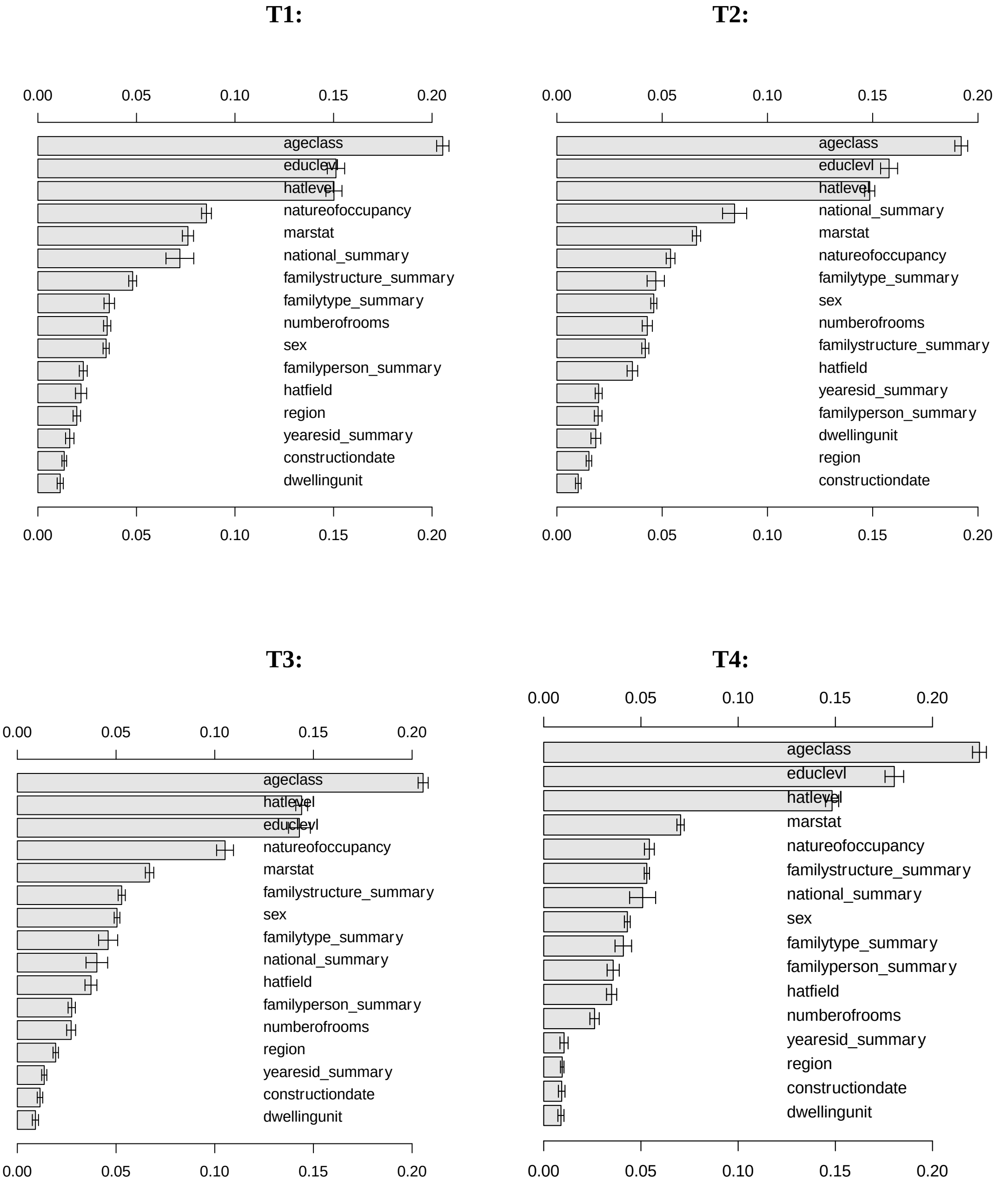
Variable selection and model performance:

Estimating significance of the predictor variables for a model allows to use the most significant ones for the training, which eventually leads to a better fit and improved classification performance.

In order to rank variable significance, we used the sensitivity analysis method proposed by Kewley et al. [26], 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

$$S_g = \sum_{j=2}^{L} |\hat{y}_{a_j} - \hat{y}_{a_{j-1}}|/(\text{L-1}) \qquad (16)$$

using the gradient measure. Figure 6 shows the significance of all variables for each term T1 to T4. It is evident, that the top three factors determining employment status are AGECLASS, HATLEVEL and EDUCLEVEL, followed by NATUREOFOCCUPANCY, NATIONAL_SUMMARY, and MARSTAT. Rank of the latter three depends on the term considered. Results also show factors with least importance, among which are YEARESID_SUMMARY, REGION, FAMILYPERSON_SUMMARY, DWELLINGUNIT, and CONSTRUCTIONDATE.


T1:

T2:

T3:

T4:

Figure 6: Variable significance using gradient measure of sensitive analysis for the periods T1-T4.

Aiming to improve the models, we considered both variable significance and correlation analysis to explore how elimination of variables affects model performance. Candidates for elimination were variables with low significance and those with high correlation to other variables. We approached to the elimination process using a variation of the backward elimination strategy, starting from the full set of 16 variables, then eliminating candidates one by one consecutively. If an elimination leads to improvement, it was used in the next step, where another candidate was eliminated, etc. The process ends when further improvement is impossible. Results are illustrated in Figure 7. Striped pattern bar shows model performance with all variables, which is the base case; light grey bars represents variable subsets underperforming the base case; dark grey bars show subsets outperforming the base one; the black bar shows the best variable selection achieved by elimination of three variables: HATFIELD, FAMILYPERSON_SUMMARY, and CONSTRUCTIONDATE.

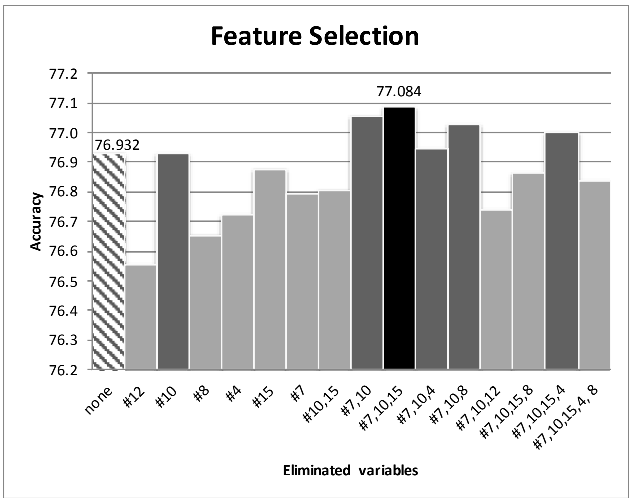


Figure 7: Feature selection by backward elimination. Max accuracy achieved by eliminating CONSTRUCTIONDATE HATFIELD, and FAMILYPERSON_SUMMARY.

Variable analysis:

In order to analyse how individual variables contribute to employment, we did variable effect characteristic (VEC) analysis [27] of the three most significant variables EDUCLEVEL, HATLEVEL, and AGECLASS, which consistently appear as top three no matter of the term considered. We also explored the next in rank variables, which vary over the terms. Being part of the sensitivity analysis, VEC explores variables role in the classification task by plotting their values x_{a_j} (x-axis) versus the responses $\hat{y}_{a_j}$ (y-axis). Interpreting VEC curves over the time T1 – T4 allows to interpret employment factors in the context of improving economic climate during the recovery since the 2012 downturn.

Figure 8 shows that, age class of the respondents (AGECLASS) is the top factor for employment over the entire period T1-T4. It indicates that people of class 8, which corresponds to age 35 - 40, are mostly employed. Youngsters of age 15 to 19 represented by class 4 have moderate employment. After class 4 employment gradually grows, reaching its peak at class 8. Also in T1 youngsters' employment value is below 6, whereas in term T4 it is above 6. The implication is that over economic recovery youngsters' employment increases as well. After the peak at class 8, employment level goes down steadily hitting its minimum at class 14, which corresponds to age 65-69. That observation implies, that after mid age, people's employment rate steadily decreases no matter of the period considered. The AGECLASS graphs have minor deviation in results.

Next in rank factor is education level (EDUCLEVEL) or equivalent training, in which respondents have been involved during the last 4 weeks (formal education). Figure 8 shows that unemployment is highest at education levels 1 to 4, which corresponds to primary education (1), lower-secondary (2), upper-secondary (3), and post-secondary non-tertiary (4). Employment slightly increases at short-cycle third level education (5) corresponding to diplomas, and dramatically increases with Bachelor's degree (6) and above, hitting the top with PhD degree (8). An interesting observation is that in T3 term, the left-hand tail of the graph is lifted above the usual, showing that primary level educated have higher employment than secondary level people. A plausible explanation of that observation is that rapidly developing sectors, such as building and construction, offer many low education job positions.

Highest level of education successfully completed (HATLEVEL) is a variable similar to EDUCLEVEL, but it represents completed education. Relationships between employment and education illustrated by the HATLEVEL graphs show that completed third level education attains highest employment.

Nature of occupancy of dwelling (NATUREOFOCCUPANCY) appear as next in rank factor for employment for the SVM model. Arguably, owning house can be seen as result of employment, not as factor for employment, although the opposite also makes sense. The housing market, house prices, and mortgage market also play significant role in the nature of occupancy values. Graphs of the nature of occupancy in Figure 8 show that house owners (1) have highest employment level, perhaps because buying a house via mortgage from credit institutions requires strong employment record. Houses purchased from local authorities under social schemes (2) indicates lower incomes, typically associated with other employment factors, such as education, thus showing lower employment rate. Employment of that category of occupants, however, increases over the time along with the economic recovery starting from T1 employment rate 0.53 to the T4 rate 0.73. Respondents who are rentals of unfurnished (4), partly-furnished (5), or furnished (6) houses have lowest level of employment, but it increases over the time from 0.55 for T1 to 0.68 for T4. The right-hand tail of the graphs represents rent-free occupants who are not owners, most likely they are family members. In the beginning of the period (T1 and T2) their employment rate is high, but in terms T3 and T4 employment drops down significantly.

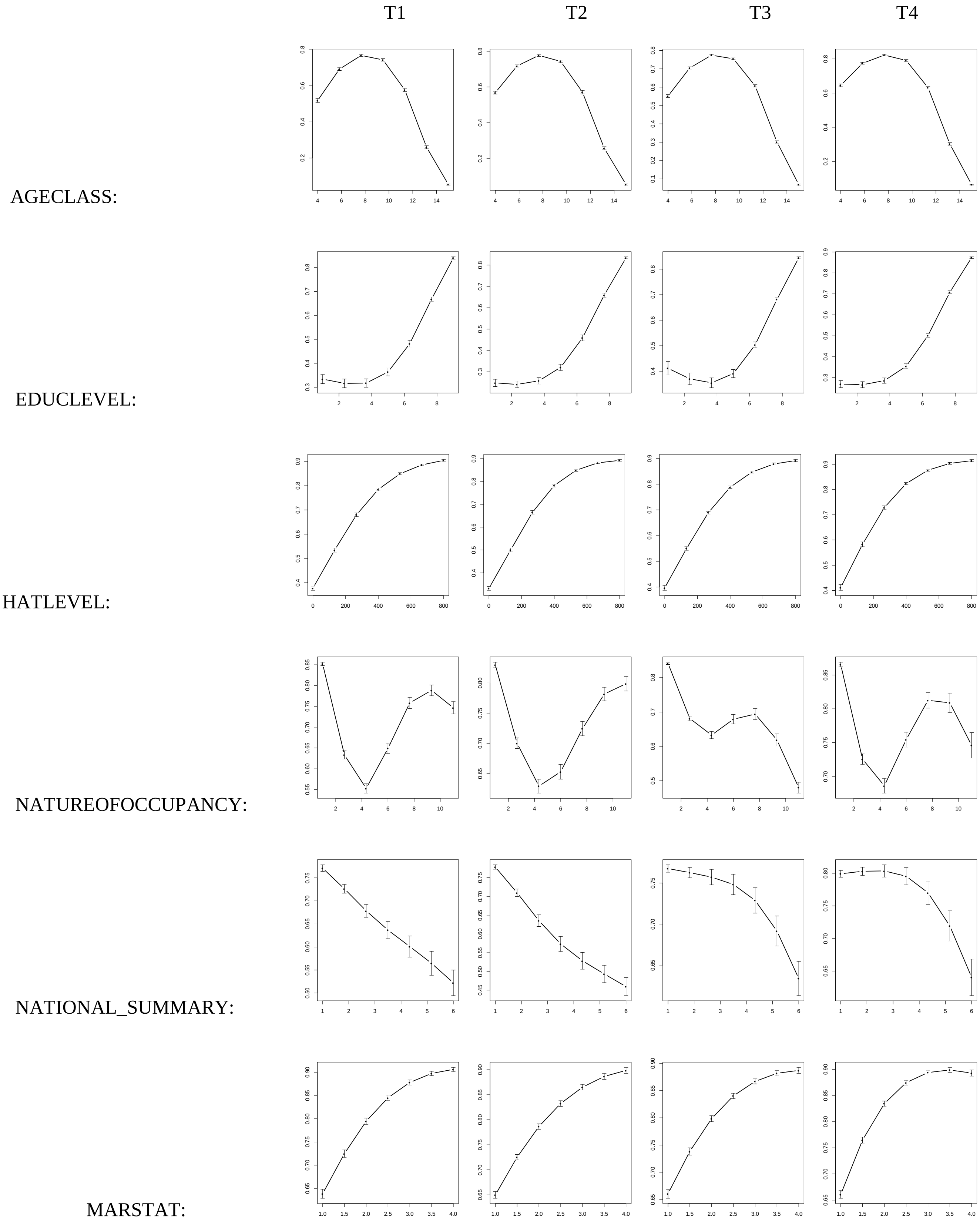


Figure 8: Vertically averaged VEC curves (points and whiskers) for AGECLASS, HATLEVEL, EDUCLEVEL, NATUREOFOCCUPANCY, NATIONAL_SUMMARY, and MARSTAT for terms T1-T4.

Arguably, an explanation of this observation is that over the economic recovery and expanding housing market, employed family members leave family houses becoming new house owners.

Nationality of the respondents (NATIONAL_SUMMARY) is another employment factor for the SVM model. It counts employment of the following nationalities: Irish (1), UK (2), EU15 (3), other EU (4), North America (5), rest of world (6). Figure 8 shows that in the beginning of the period T1 unemployment among foreign nationals is higher than at the end of the period in term T4, when locals and EU nationals have nearly the same employment rate.

Finally, the SVM model determines the marital status (MARSTAT) of the respondent as employment factor. Status values are: single (1), married (2), widowed (3), and divorced or separated (4). Figure 8 shows that singles, usually young people have lowest employment rate, perhaps due to other factors, such as age and education. Next is the category of married, where unemployment level is caused by maternity or childcare. The other categories have highest employment as those factors are not issue.

In summary, using SVM models and VEC curves as tools for analysis of predictor variables allows for measurement and interpretation of employment factors and tracking down their change over the time.

CONCLUSION

Objective of this study is to analyse data from Quarterly National Household Survey (QNHS), a large-scale nation-wide survey, by the means of SVM. It addresses some gaps in previous research, focusing to measuring factors and providing insight with regard to their role in employment. We explored experimentally how choice of SVM kernel and parameter values affect the model and built the best performing one. Results show that Gaussian RBF kernel with sigma=0.051 and C=3.5 are best for that dataset. Performance was measured by accuracy of prediction and AUC from ROC analysis.

Analyzing variable importance for the model, we estimated the factors that affect the employment status of respondents. After ranking those factors, we focused to the most significant ones: age class, current and completed education, nature of occupancy, nationality, and marital status. Further detail for each factor was given by VEC analysis, which reveals how values of those factors contribute to the employments status.

In conclusion, we find SVM as a powerful data mining tool, that allows for analysing raw data in order to obtain valuable insights with practical meaning in the employment domain.

REFERENCES

[1] Dash, M. and Singhania, A., 2009, “Mining in Large Noisy Domains”, J. Data and Information Quality, 1, pp. 1-30.

[2] Salfner, F., Lenk, M. & Malek, M., 2010, “A survey of online failure prediction methods. “, ACM Comput. Surv., 42, pp. 1-42.

[3] Cortes, C. and Vapnik, V., 1995, “Support-vector networks.”, Machine Learning, 20(3), pp. 273-297.

[4] Drucker, H. Burges, C., Kaufman, L., Smola, A., and Vapnik, V., 1997, “Support vector regression machines”, Advances in Neural Information Processing Systems 9, pp. 155-161.

[5] Oladunni, O. O. & Singhal, G., 2009, “Piecewise multi-classification support vector machines.”, International Joint Conference on Neural Networks, IJCNN'09.

[6] Burges, C., 1998, “A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery”, 2, pp.121-167.

[7] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J., 2009, “Modeling wine preferences by data mining from physicochemical properties.”, Decision Support Systems, 47, pp. 547-553.

[8] Horng, S., Su, M., Chen, Y., Kao, T., Chen, R., Lai, J. and Perkasa, C., 2010, “A novel intrusion detection system based on hierarchical clustering and support vector machines.”, Expert Systems with Applications, 38, pp. 306-313.

[9] Fei, L., Li, W. & Yong, H., 2008, “Application of least squares support vector machines for discrimination of red wine using visible and near infrared spectroscopy.”, Intelligent System and Knowledge Engineering, ISKE' 08.

[10] Wang, X., Lv, J. & Xie, D., 2010, “A hybrid approach of support vector machine with particle swarm optimization for water quality prediction.”, 5th International Conference on Computer Science and Education (ICCSE), pp. 1158–1163.

[11] Selvanayaki, M., Vijaya, M. S., Jamuna, K. S. and Karpagavalli, S., 2010, “An Interactive Tool for Yarn Strength Prediction Using Support Vector Regression. “, Conference on Machine Learning and Computing (ICMLC), pp. 335-339

[12] Petrujkic, M., Rapaic, M. R., Jakovljevic, B. & Dapic, V., 2008, “Electric energy forecasting in crude oil processing using Support Vector Machines and Particle Swarm Optimization.”, 9th Symposium on Neural Network Applications in Electrical Engineering, NEUREL.

[13] CSO: QNHS, [Online], http://www.cso.ie/en/qnhs/, 2016.

[14] R Development Core Team, 2009, “R: A language and environment for statistical computing.”, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

[15] Cortez, P., 2010, “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, pp. 572– 583.

[16] Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005, “ROCR: visualizing classifier performance in R.”, Bioinformatics 21(20), pp. 3940-3941.

[17] Fawcett, T., 2005, “An introduction to ROC analysis”, Pattern Recognition Letters 27(8), pp. 861–874.

[18] Kelly, E., McGuinness, S., 2014, “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

[19] 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, pp. 616-634.

[20] Jantavan, B., Tsai, C., 2013, "The Application of Data Mining to Build Classification Model for Predicting Graduate Employment", International Journal of Computer Science and Information Security, 11(10).

[21] Mishra, T., Kumar, D., 2016, "Students' Employability Prediction Model through Data Mining", International Journal of Applied Engineering Research, 11(4), pp. 2275-2282.

[22] Sapaat, M., Mustapha, A., Ahmad, J., Chamili, K., Muhamad, R., 2011, "A Classification-based Graduates Employability Model for Tracer Study by MOHE", Digital Information Processing and Communications, Springer Berlin Heidelberg, pp. 277-287.

[23] Alsultanny, Y., 2013, “Labor Market Forecasting by Using Data Mining”, International Conference on Computational Science, Procedia Computer Science 18, Elsevier, 2013, pp.1700-1709.

[24] Kirimi, J., Moturi, C., 2016, "Application of Data Mining Classification in Employee Performance Prediction", International Journal of Computer Applications, 146(7), pp. 28-35.

[25] Xiaoh, W., 2009, “Intelligent Modeling and Predicting Surface Roughness in End Milling. “, Fifth International Conference on Natural Computation, ICNC '09.

[26] Kewley, R., Embrechts, M., Breneman, C., 2000, “Data strip mining for the virtual design of pharmaceuticals with neural networks.”, IEEE Transactions on Neural Networks, 11 (3), pp. 668–679.

[27] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009, “Modeling wine preferences by data mining from physicochemical properties”, Decision Support Systems, 47(4), pp. 547–553.

[28] Nachev, A., 2017, “Using multi-layer perceptrons for analysis of labor data”, In Proc. of International Conference Artificial Intelligence, ICAI’17, Las Vegas, 17-20 Jul, pp.223-229.