International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-10 Issue-3, February 2021

Analysis of Unmet Healthcare Needs in Ireland: A Data Mining Approach

Anatoli Nachev

Abstract: ***This study explores data form the Survey of Income and Living Condition (SILC), related to factors contributing to unmet healthcare needs in Ireland. We analysed predisposing, enabling and needs factors by building predictive models and measured the predictor importance by sensitivity analysis. Results show that critical factors for meeting the healthcare needs include financial status, degree of urbanization, indicators of social exclusion and deprivations, and self-perceived general health condition. Identifying and quantifying those factors form raw data may facilitate decision making in the domain.***

Keywords**: unmet healthcare, data mining, classification, logistic regression.**

I. INTRODUCTION

Objective and contribution of this study is to explore the latest Irish data on income and leaving condition in order to identify factors affecting the unmet healthcare needs in the country and how they impact the healthcare system in place. Data also contains new variables, not mentioned in research publications before, so this study enlarges the scope of factor analysed and provides an empirically proven conclusions that can be used for decision making in the domain.

A. Unmet healthcare needs

Unmet health care needs can be defined as a measure of access to the healthcare system, which describes situations in which someone who needed healthcare did not receive it. It can also be defined as the difference between the healthcare services considered as necessary for a specific health problem and the services received [6]. Unmet need for healthcare can also be seen as covering a spectrum of healthcare needs that are not optimally met. At one end there is unexpressed demand - people who have healthcare needs but who are not aware of them, or who choose not to seek healthcare. At the other end there is expressed demand that is sub-optimally met. This can include people ineligible for treatment, or who have poorer quality treatment than would optimally be the case. For some individuals, their unmet need may be a combination of the two. The unmet healthcare needs can play important role as indicators for measuring inequalities and imbalances in access to the healthcare at the national and local levels [7].

Manuscript received on February 08, 2021.
Revised Manuscript received on February 15, 2021.
Manuscript published on February 28, 2021
* Correspondence Author
Dr. Anatoli Nachev*, Lecturer, Business Information Systems, J. E. Cairnes School of Business & Economics, National University of Ireland, Galway, Ireland.

II. MATERIALS AND METHODS

A. Data mining approach

Data mining is a process that requires methodology for carrying out projects successfully and for reaching the project objectives set out at the beginning. In order to achieve the initial objectives, the process needs to translate them into data mining tasks, to set success criteria, to understand data and their values, to apply appropriate data transformations, to develop test strategy, select modelling techniques, and provide means for evaluating the model performance and analysis of results. The CRISP-DM (CRoss-Industry Standard Process for Data Mining) [1] is an industry-proven standard methodology and framework for running data mining projects. It is also widely adopted by researchers and practitioners who process raw data in order to train and build predictive models and analyse results. This study uses the CRISP-DM methodology as research approach because exploring factors of unmet healthcare needs is made by training a predictive model using the SILC data and identifying factors that contribute to a successful prediction. CRISP-DM has six stages in a cycle (Fig. 1) and their role is briefly outlined below.

Business understanding stage determines the project objectives and translates them into data mining task and sub-tasks. It also sets success criteria from business and data mining point of view.

Data understanding stage deals with data gathering and understanding, provides initial statistical analysis and visualization, where applicable.

Data preparation stage consists of activities related to data elimination, constructing single dataset from the multiple data sources, and data pre-processing.

The modeling stage considers appropriate modeling techniques, deals with data partitioning and test strategies, and builds predictive models. It also evaluates the model performance and analyses results.

At the evaluation stage, models are checked whether they meet the success criteria and review the data mining process for possible improvements.

The deployment stage requires the model implementation in business environment and also developing monitoring and maintenance plans. [2]

With reference to the business understanding stage, our data mining task is binary classification as the dependent variable unmet healthcare needs is binary (yes/no). This also requires analysis of results in order to estimate the factors that affect the unmet healthcare needs in Ireland.

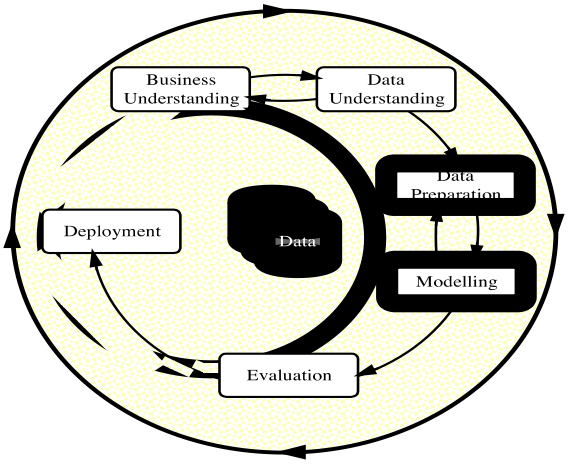

Figure 1. Phases of the CRISP-DM process model for data mining.

With reference to the data understanding stage, data were acquired and processed as outlined below.

B. Data and sample design

This research studies cross-sectional micro-data collected by the Survey on Income and Living Conditions (SILC) carried out in Ireland in 2018 (released 2020). Data were obtained from the EUROSTAT EU-SILC data as part of project RPP228/2018-EU-SILC and met criteria for access to confidential data [2]. Data are based on a nationally representative probability sample of the population residing in private households within the country. Some variables use household as unit of measure while other are measured at person level. For Ireland, cross-sectional data represent a minimum of 3750 households and 8000 persons aged 16 or over to be interviewed. There are missing data caused by non-response, missing household members, or unavailable information. The Irish EU-SILC data are stored in four files:

- Household Register (D-file) contains common household information on weights, sampling, regional identifiers and degree of urbanization. Contains only information at the household level in 17 variables and 5030 observations.
- Household Data (H-file) contains specific socio-demographic household information such as income, economic situation, poverty and employment indicators. Data are presented in 230 variables and 5030 observations.
- Personal Register (R-file) contains mostly identifiers about family relations, basic demographic information, and childcare usage at personal level. It is organised in 58 variables and 12,617 observations, some of which represent children (age below 16).
- Personal data (P-file) contains socio-demographics information at personal level, such as income, work, unemployment, health, nationality, migration, work intensity, etc. It contains 255 variables and 9804 observations.

Totally, data contains 560 variables, most of which are not relevant to the required analysis and data mining task and were therefore eliminated with reference to the CRISP-DM's data understanding and data preparation stages. Household and personal data were also integrated into a single dataset, on the basis of matching household ID in D and H files and person's household ID in R and P files. The combined dataset contains 28 variables and 9589 observations.

C. Variables

According to the Andersen's behavioural model of health services [8], predictor variables for unmet health care needs can be grouped into predisposing factors, factors that enable the use of health care, and factors that indicate the need for using health care services. The dataset variables were placed in those three groups as follows:

- *Predisposing factors*, such as age and gender, represent biological imperatives that suggest using the health services. Socio-demographic factors such as education, marital status, current economic status, and degree of urbanization can also be seen as predisposing for unmet needs for healthcare services. Health beliefs, attitudes, and knowledge that people have about health and healthcare can also be described as predisposing. Variables, such as country of origin and citizenship can capture some of the cultural differences on perception of the health services.
- *Enabling factors* for using the healthcare services represent personal, family, and social resources that can enable access to the services or being a barrier to them. Variables of this category can be household disposable income; ability to afford paying for one-week annual holiday away from home; capacity to afford a meal with meat, chicken, fish (or vegetarian equivalent) every second day; capacity to face unexpected financial expenses; ability to make ends meet; financial burden of medical care; financial burden of medicines; ability to spend a small amount of money each week on yourself; and gross monthly earnings for employees.
- *Needs factors* indicate the need for health services, described by variables for self-perceived health status, such as general health; presence of chronic illness or condition; limitation in activities because of health problems; number of consultations of a general practitioner or family doctor; number of consultations of a medical or surgical specialist; body-mass-index (BMI); type of physical activity when working; time spent on physical activities (excluding working) in a typical week; frequency of eating fruits and vegetables.

The dependent variable, which represents unmet healthcare needs is based on two questions: "Was there any time during the past 12 months when you really needed medical examination or treatment for yourself?", "Did you have a medical examination or treatment each time you really needed?". Answer values are 1-yes, there was at least one occasion when the person really needed examination or treatment but did not receive it; 2-no, there was no occasion when the person really needed examination or treatment but did not receive it.

Table 1 summarizes the variables and shows some basic statistics.

Table I: Variables and basic statistics.

variables	min	max	mean	median	sd	mad	skew	kurtosis	se
Predisposing factors									
Age	1	4	2.660	3	1.009	1.4826	-0.018	-1.164	0.010
Sex	1	2	1.521	2	0.500	0	-0.084	-1.993	0.005
EducLevel	0	50	5.663	0	14.919	0	2.349	3.747	0.152
HighestEduc	100	500	338.295	344	150.186	231.2856	-0.294	-1.350	1.534
MarStat	1	5	1.945	2	0.957	0	1.396	1.873	0.010
Urban	1	3	1.960	2	0.892	1.4826	0.079	-1.738	0.009
Country	1	3	1.199	1	0.497	0	2.497	5.332	0.005
Citizen	1	3	1.107	1	0.359	0	3.565	12.796	0.004
EconStatus	1	11	4.630	5	3.199	4.4478	0.261	-1.304	0.033
EarningsMonth	0	30438.33	1228.847	0	2115.075	0	2.657	11.909	21.599
Enabling factors									
PayWeekHoliday	0	2	1.335	1	0.476	0	0.650	-1.441	0.005
PayMeal	0	2	1.015	1	0.125	0	7.562	58.614	0.001
PayUnexpected	0	2	1.378	1	0.494	0	0.397	-1.565	0.005
MakeEndsMeet	0	6	3.289	3	1.172	1.4826	-0.008	-0.105	0.012
FinBurden	0	3	2.087	3	1.172	0	-0.892	-0.804	0.012
FinBurdenMed	0	3	2.145	3	1.118	0	-1.011	-0.468	0.011
SpendWeekly	0	3	0.768	1	0.588	0	0.363	0.832	0.006
EatingFruit	1	6	2.184	2	1.312	1.4826	1.236	0.801	0.013
EatingVeges	1	6	1.954	2	0.990	0	1.647	3.534	0.010
Needs factors									
GeneralHealth	1	5	1.810	2	0.842	1.4826	0.909	0.552	0.009
ChronicIllness	1	2	1.697	2	0.460	0	-0.856	-1.267	0.005
LimitActiv	1	3	2.740	3	0.570	0	-2.089	3.156	0.006
ConsultGP	0	5	2.446	2	1.178	1.4826	0.625	-0.264	0.012
ConsultSpecialist	0	5	1.479	1	0.825	0	1.999	4.368	0.008
BMI	0	40	19.239	24	11.621	5.9304	-0.817	-0.794	0.119
PhysicalActivityWork	0	4	1.289	1	1.331	1.4826	0.588	-1.037	0.014
Other									
ReasonUnmet	0	8	0.072	0	0.504	0	10.365	130.547	0.005
Dependent									
UnmetNeeds	0	1	0.968	1	0.176	0	-5.316	26.261	0.002

where *sd* is standard deviation, *mad* is median absolute deviation (from the median), and *se* is standard error.

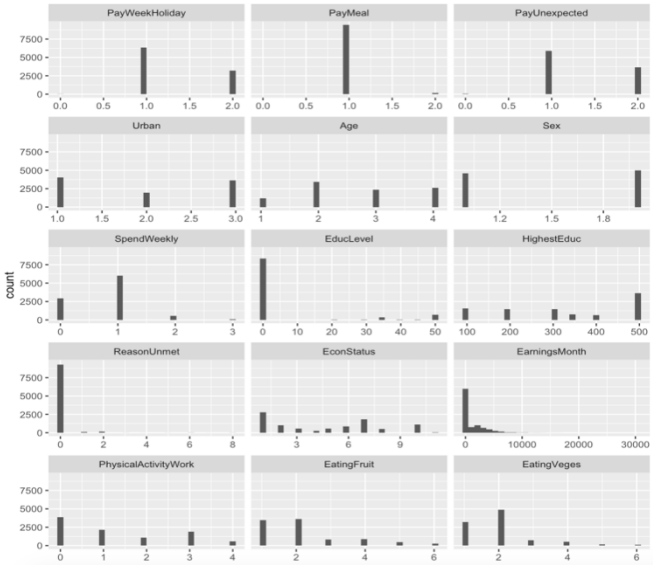

Fig 2 also shows the variables histograms.

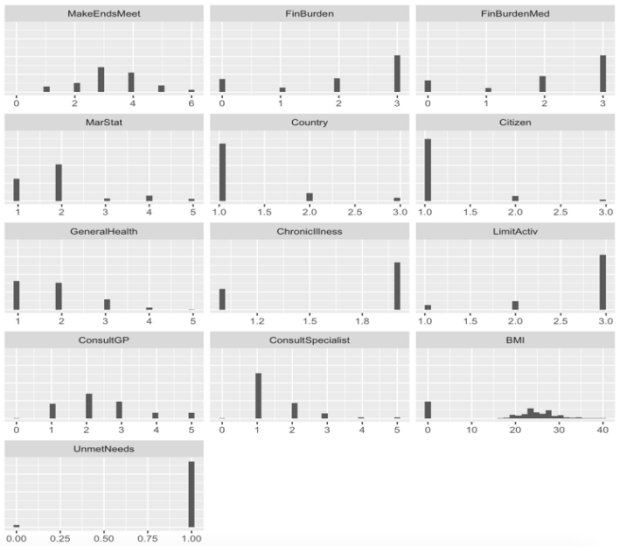

Figure 2. Variable histograms.

D. Modelling Techniques

In relation to the modelling stage of CRISP-DM, the technique selected to build predictive model is logistic regression [12], [13]. It is a statistical method that measures relationship between one or more independent variables X_i and categorical dependent variable Y by estimating probabilities using the logistic function

$$Y(X_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_{1n})}} \quad (1)$$

also known as odds ratio, where β_i are regression coefficients. The logistic function will always produce an S-shaped curve as shown in Fig 3, so values of Y close to 0 indicate very low probability of belonging to the success class 1 and values close to 1 indicate high probability of belonging to the success class 1.

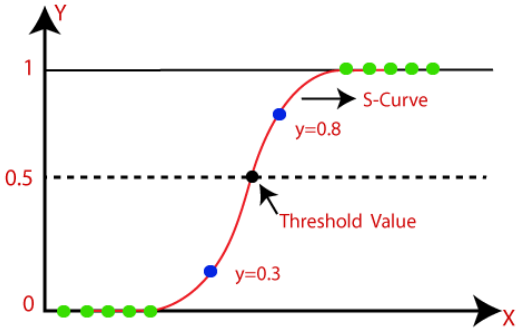

Figure 3. Logistic regression function.

The coefficients β_i are estimated using the maximum likelihood statistical technique on the training data.

E. Performance Estimation

The most common approach for estimating performance of binary classifiers is to use confusion matrix containing four categories of responses: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The model accuracy, computed as:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Accuracy is primary measure for model performance, but that metric can be unreliable when the dependent variable contains unbalanced class label distribution, which is the case of our dataset. The Receiver Operating Characteristics (ROC) analysis [14] and its scaler AUC address the Acc disadvantage, providing another way to measure the model performance. ROC curves show relation between sensitivity and specificity of the model over the entire range of operation points from 0 to 1. The model sensitivity is represented by the true positive rate (TPR) and the model specificity by the false positive rate (FPR). AUC is the area under the ROC curve, which is ranging from 0 to 1. The greater it is, the better performance we have, regardless of the choice of operation points - something that Acc cannot offer.

III. RESULTS AND DISCUSSION

Data were processed and models built using R environment [9]-[11], exploring the three categories of factors as follows:

A. Predisposing factors.

We built a logistic regression model using the relevant variables and also stratified sampling from randomly selected partition comprising 80% of the original data. The rest of 20% data were used for testing only. The model shows 97.080% accuracy. This, however can't be credible performance metric due to the fact that the 'yes' class label is over-presented (96.72% of all samples). As mentioned above, we did ROC analysis. The ROC curve in Fig. 6 shows the relationship between the model sensitivity and specificity and AUC=0.617, which is indication of fair performance.

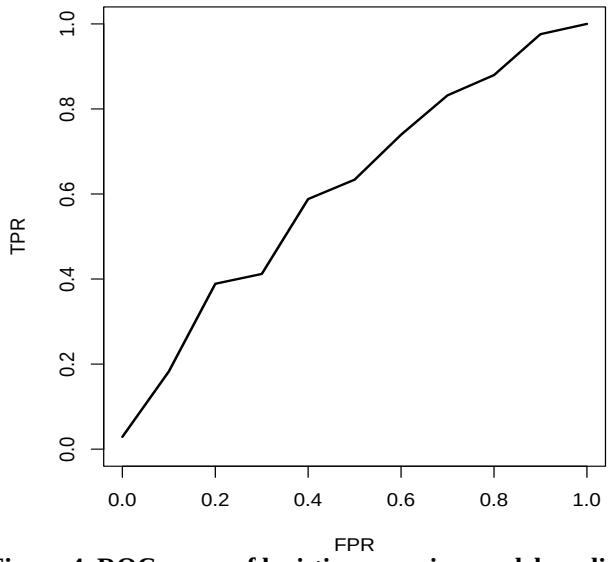

Figure 4. ROC curve of logistic regression model applied to predisposing data. ACC= 97.080, AUC= 0.617

Further sensitivity analysis of the variables shows their importance and contribution to distinguishing between met and unmet healthcare needs. Fig 7 ranks the variables according to their importance in average, which was achieved after 10 runs of that analysis. Whiskers show the importance variance in those runs.

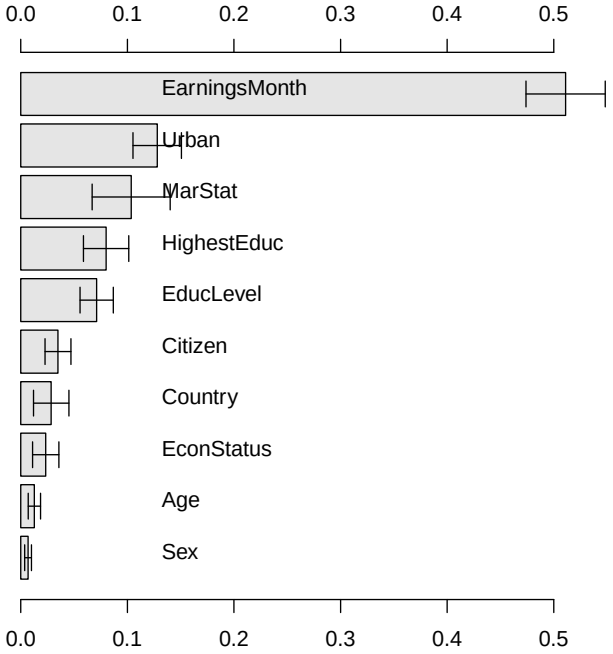

Figure 5. Predisposing factor importance using sensitivity analysis

Results show that the variable gross monthly earnings for employees is highest in importance among the predisposing factors. That variable represents the monthly amount of earnings for employees and shows their ability to meet costs of services and facilities that require payment. The Irish health system provides higher quality services as part of various voluntary health insurance plans, which is captured as one of the primary factors for meeting the healthcare needs. Next in importance is the factor degree of urbanisation with three values: densely-populated area, intermediate area, and thinly-populated area. Being a socio-demographic factor, it shows that individuals living in highly populated areas such as cities, have better access to the health facilities and services than those living in rural areas.

The lowest in importance factors according to results are sex and age. The implication of that finding is that the gender and age do not play major role in meeting the health needs in Ireland and therefore individuals are treated equally in relation to those criteria.

B. Enabling factors.

Similarly, a logistic regression model was built using the enabling variables and stratified sampling of 80% partition of the original data and 20% data used for testing. The model shows 96.715% accuracy and the ROC curve shown in Fig. 8 with a higher AUC=0.728. This is indication of better performance and therefore according to the results, enabling factors play more important role in determining the unmet healthcare needs.

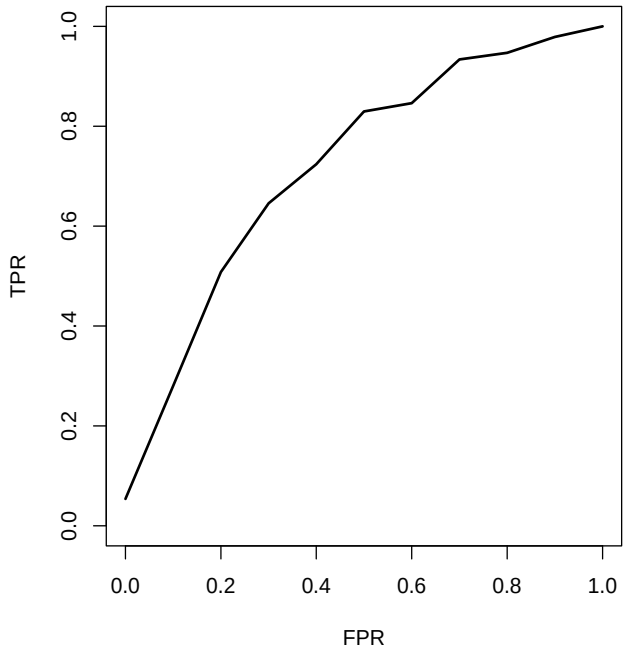

Figure 6. ROC curve of logistic regression model applied to enabling data. ACC=96.715, AUC= 0.728

Similarly, the sensitivity analysis of the variables shows their role in distinguishing between the two classes Ranking after 10 runs is illustrated in Fig 9. Whiskers show the variance in those runs.

Retrieval Number: 100.1/ijeat.C22320210321
DOI:10.35940/ijeat.C2232.0210321
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

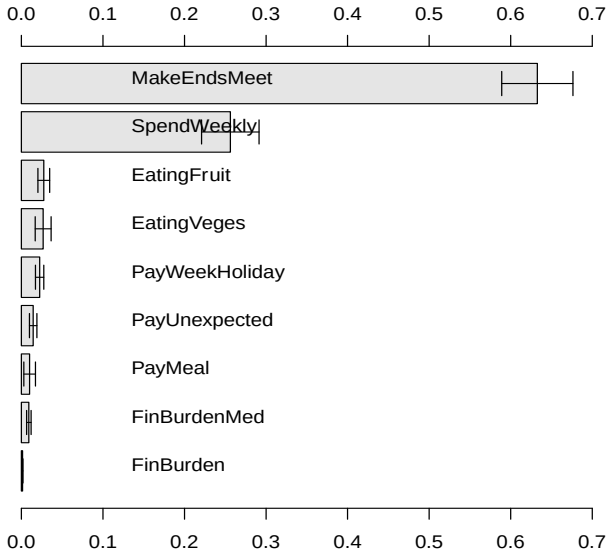

Figure 7. Enabling factor importance using sensitivity analysis

Results imply that the factor ability to make ends meet is most important as indicator of social exclusion and deprivations. Values range from difficult to easy.

Next in importance factor is ability to spend a small amount of money each week on yourself. It is indicator of financial status and ability to afford paying voluntary health insurance policy. Lowest in ranking is the factor perception of the extent to which costs for medicines (prescribed and non-prescribed) are a financial burden to the household. This is indication that cost for medicine is not big issue for the Irish citizen as the system for providing social benefits for covering expenses above certain limit does not make this factor important.

C. Needs factors.

A logistic regression model was built using the needs variables. The data used for training were 80% of the original one as before and 20% were used for testing. The model shows 96.194% accuracy and the ROC curve in Fig 8 shows AUC=0.711. This is indication that overall needs factors are more important than predisposing, but less important that the enabling one.

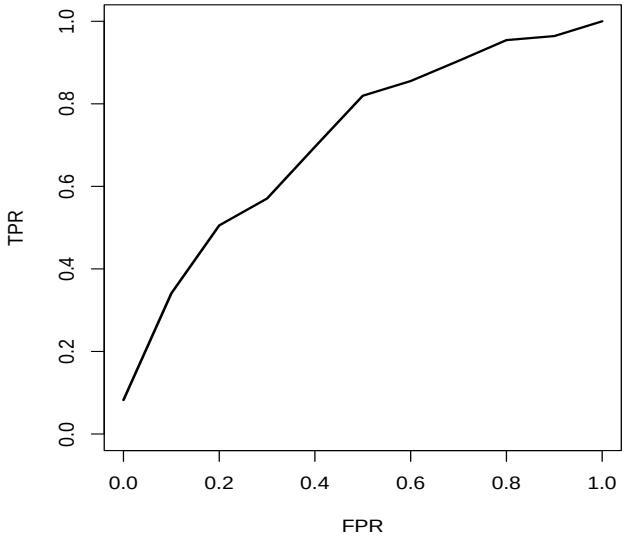

Figure 8. ROC curve of logistic regression model applied to needs data. ACC=96.715, AUC= 0.728

Similarly, the sensitivity analysis of the variables shows their role in distinguishing between the two classes Fig. 9 shows their ranking on average after 10 runs. As it stands, the only factor that determines meeting or not meeting the healthcare need is the self-perceived general health condition with values ranging from very bad to very good. Despite being subjective, that factor reflects different dimensions of health, and symptoms, and is therefore related strongly to the self-perceived estimation of whether healthcare needs have been met or not. Among factors playing minor role of meeting the healthcare needs are BMI, and presence of chronic illness or condition, both suggesting that the healthcare system covers adequately that kind of conditions and therefore the patients don't experience problems with respect to receiving treatment based on those conditions.

Another factor not affecting significantly meeting the healthcare needs is the number of consultations of a medical or surgical specialist, suggesting that the access to those specialists is adequate, given such consultations are recommended by the GPs.

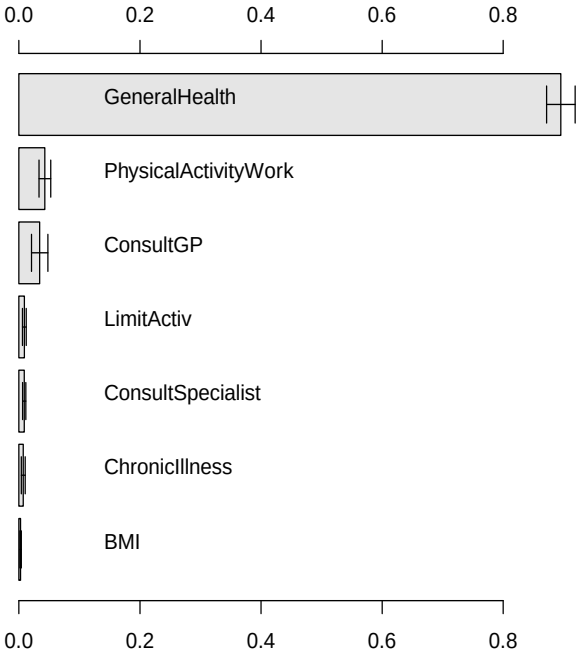

Figure 9. Needs factor importance using sensitivity analysis

IV. CONCLUSION

Objective of this study is to analyse data form recent survey on income and living conditions, particularly focusing to factors revealing how healthcare needs are met by the Irish healthcare system. In addition to previous research in the area, this study addresses more factors arising from the recent data updates and also applying different approach, which estimates the factors by building predictive models and applying sensitivity analysis to measure importance of those factors. The factors were grouped into three categories: predisposing, enabling, and needs, and separate models were built for each of these groups. Factors importance was analysed within each group. Experimental results show that gross monthly earnings for employees is the most important predisposing factor.

Among the enabling factors, indicators of social exclusion and deprivations play major role, whereas self-perceived general health condition is the primary indicator for meeting healthcare needs or not. Experiments also quantify the factors, which may facilitate decision making in the subject domain.

REFERENCES

1. 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.
2. European Commission. *EU-SILC User Database Description*. Luxembourg: European Commission, Nov, 2019.
3. S. Connolly, S. and Wren, M. "Universal Health Care in Ireland—What Are the Prospects for Reform?", Health Systems & Reform, 5:2, 94-99, 2019.
4. S. Connolly, S. and Wren, M. "Unmet healthcare needs in Ireland: Analysis using the EU-SILC survey", Health Policy, vol.121, 434–441, 2017.
5. Popovic N., Terzic-Supic Z., Simic S., Mladenovic B. "Predictors of unmet health care needs in Serbia; Analysis based on EU-SILC data.", PLoS ONE 12(11), 2017.
6. Carr W., Wolfe S. "Unmet Needs as Sociomedical Indicators.", International Journal of Health Services.", 6 (3):417–430, 1976.
7. Allin, S., Hernández-Quevedo, C., Masseria, C., "Measuring equity of access to health care.", In Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects., pp 187-221, Cambridge University Press, 2009.
8. Andersen R. "Revisiting the Behavioral Model and Access to Medical Care: Does it Matter?", Journal of Health and Social Behavior, 36:1, 1995.
9. 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.
10. 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.
11. Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., "ROCR: visualizing classifier performance in R.", Bioinformatics 21(20), pp. 3940-3941, 2005.
12. Menard, S. *Applied Logistic Regression* (2nd ed.). SAGE, 2002
13. Hyeoun-Ae Park, 2013, "An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain", Journal of Korean Academy of Nursing 43(2):154-164.
14. Fawcett, T., 2005, "An introduction to ROC analysis", Pattern Recognition Letters 27(8), pp. 861–874.

AUTHORS PROFILE

Dr. Anatoli Nachev received his PhD degree in BAS, Institute of Math and Informatics, section AI. He received his MSc and BSc degrees in SU, FMI. He is currently a lecturer at Business Information Systems, J. E. Cairnes School of Business & Economics, National University of Ireland, Galway, Ireland. Research interests include business intelligence and predictive modeling, machine learning, data mining, Artificial Intelligence, etc. He has numerous publications in books, international journals and conferences in the fields of interest.