Usage of Information and Communication Technologies in Italian Enterprises

Fabio Caironi

May 12, 2021

Course:	Statistical Learning, Deep Learning and Artificial Intelligence
A.A.:	2019-2020
Instructor:	Professor S. Salini

Abstract

Italian enterprises are riding the wave of world technological advancement and are day by day converting their ICT infrastructures to keep up with the latest services and utilities. In this lab project, I've inspected the Istat survey of 2018 about Italian enterprises in order to gain insight into which factors contribute to the technological development of an enterprise. In particular, I've looked into the usage of cloud computing services and its determinants and tried to give an overview on the results of the survey by also finding similarities among the enterprises' characteristics. The statistical method used to perform the supervised learning task of predicting the probability of usage of cloud computing services was the logistic regression. The results depict the interesting framework where most IT-related variables, like internet connection speed or usage of an enterprise website, positively influence the probability of activating cloud computing services, while other indicators such as an enterprise's revenues are not helpful to classify the usage of such services.

1 Introduction and dataset description

The Italian Statistic Institute (ISTAT) carries out on annual basis a survey on Information and Communication Technologies (ICT) usage, in compliance with EU regulations and in collaboration with Eurostat on the basis of shared methodologies among the UE member Countries. The survey provides a broad set of information about ICT usage by Italian enterprises with more than 10 persons employed and belonging to different sectors (as specified by Ateco 2007 classification). Such information mainly regard: the presence of ICT specialists

and digital skills; the type of connection and Internet usage; the sales and purchases through information networks; e-invoicing; the usage of cloud computing services; the usage of 3D printing, robotics and Big Data analytics; the determinants of digital innovation of the enterprise. In addition, the survey collects structural information about the enterprise, such as revenue and the average number of employees.
A microdata table for public use containing the results of the survey is published yearly by ISTAT and is available for registered users in the microdata section[1]. The publication includes metadata such as survey methodology, dataset description and variable information, where the latter has been attached in the appendix of this paper (see 3.5).

Our study adopts the 2018 dataset, which is 22079×224 long, with single respondent enterprises as rows and answers as columns, which are mostly categorical and indicate whether an enterprise has activated/performed a service/action/investment related to the ICT usage.
After an exploratory analysis of the dataset we will try to predict, through a supervised statistical technique, the binary variable D1 "Usage of cloud computing services on the internet", that indicates the adoption of such services in 2018 by the enterprise.

2 Exploratory Analysis

Our analysis starts from an accurate description and inspection of the dataset. We will here tackle some indispensable pre-processing steps and look at the similarity among features to preliminarily try and understand which are the factors contributing to the adoption of cloud computing services.

2.1 Numerical variables

Seven numerical variables are present in the dataset and they express percentages. Their correlation is shown in Figure 1.

High correlation is present among some variables. In particular, couples (I3a, I3b) and (I4a, I4b) are perfectly negatively correlated, being complementary percentages of sales and thus carrying the same information. We therefore drop I3b and I4b.
Also (A3_, C2_) are highly correlated (91%), where A3_ is the percentage of employees using a computer and C2_ is again the percentage of employees using a computer, but additionally with an internet connection. We could as well drop the variable A3_ since it is somehow implied by variable C2_, which looks more related to the prediction of D1.

[1] `https://www.istat.it/en/archivio/177225`

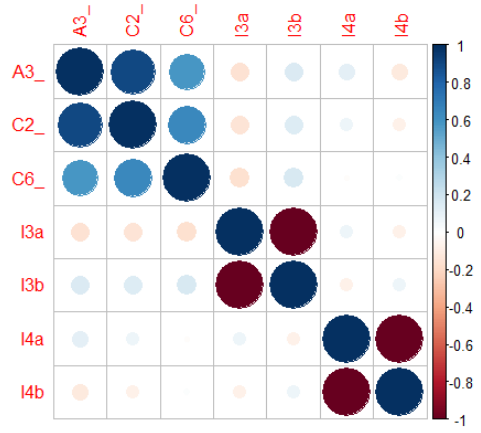

Figure 1: Correlation plot of numerical variables

2.2 NAs study

The study of missing values is performed along the two dimensions of the dataset:

- First, we recognize highly incomplete columns, that are associated to questions in the survey that most respondents couldn't answer. Such variables cannot be utilized for any meaningful analysis. It turns out that 43 out of 224 variables have 20% or more missing values and all the others have less than 5%. Here's the list of highly incomplete variables:

  B4, C6_, C7a, C7b, C7c, C9a, C9b, C9c, C9d, C9e, C9f, C11a, C11b, C11c, C11d, D2a, D2b, D2c, D2d, D2e, D2f, D2g, D3a, D3b, E2a, E2b, E2c, E2d, F2a, F2b, F2c, F2d, F2e, F2f, F2g, G2a, G2b, H2, I2_cl, I3a, I4a, I6_cl, I2I6_cl.

  The reader can verify that many of these variables correspond to specific questions in the survey about the typology of service used or other variables of the ICT services that are somehow implied by other variables in the dataset. In addition, the set D2* to D3* describe cloud computing service characteristics, that are pointless in the objective of predicting the usage of computing services itself. In order to maintain the trustworthiness of the survey, without bootstrapping or imputing the missing values, we just drop all of these 43 incomplete variables.

- Second, we look at rows with missing values: at this point, 1110 (roughly 5%) rows show at least one missing value. With the same aim as before, we drop them.

2.3 Categorical variables

The categorical variables constitute the bulk of our data and are for the most part binary, as is the answer "yes"/"no". A proper pre-processing step is usually to identify and handle collinearity in the data, before applying statistical methods – like regression – which are sensitive to high correlation among variables. With binary categorical variables, though, correlation is not defined; instead, similarity measures are suitable for capturing the likeness of two such variables. We will use the Sokal-Michener index, that counts the number of co-presences and co-absences relative to the feature length and for this reason is also called 'simple matching' similarity measure. On the one side, we will look for similarity among any pair of variables in the dataset and tag an element of each pair as useless repetition. On the other side, we will look at which variables are most similar to D1 and try to identify causality among such answers.
After the removal of constant variables C1 and C3 (they're all equal to "yes": fortunately, every enterprise of this survey has an internet connection!), we proceed with the computation of similarities.

2.3.1 Similarity among couples

Many couples are more than 50% similar (7169 couples), but we just seek *highly* similar columns, i.e. those with a S-M index of at least 0.99. Our findings are that the set of variables C12** are pairwise highly similar and the reason is simple: they come from very similar questions, exploring the same topic but in different fields of the enterprise. In order not to wrongly empower such variables as predictors in a regression model, we drop one by one the elements of each couple until any similarity above 0.99 is eliminated. The removed variables in this step are:

C12a5, C12b1, C12b2, C12b3, C12b4, C12b5, C12c5, C12d5, C12e5, C12f5, C12g5, C12h5.

2.3.2 Similarity to the target

Next, we inspect the similarity of all binary variables to the target D1. This helps us identify which variables will be able to predict D1 better, but also which ones are similar because represent, just as D2*-D3* before, a repetition of the information carried by D1. The star graph in Figure 2 shows similarity of each variable with D1.

Variables J1e1, J1e2, J1f1, J1f2 regard investments in cloud computing technologies and web applications over the past 2 years, while J2e1, J2e2, J2f1, J2f2 investments to be done in 2018 and 2019. It seems like all these variables are highly similar to D1 and after all we're not interested in finding that investments in the cloud computing sector have determined the enterprise to use cloud computing technologies, because this is obvious. As a consequence, we drop variables strictly related to cloud computing investments, namely the sets

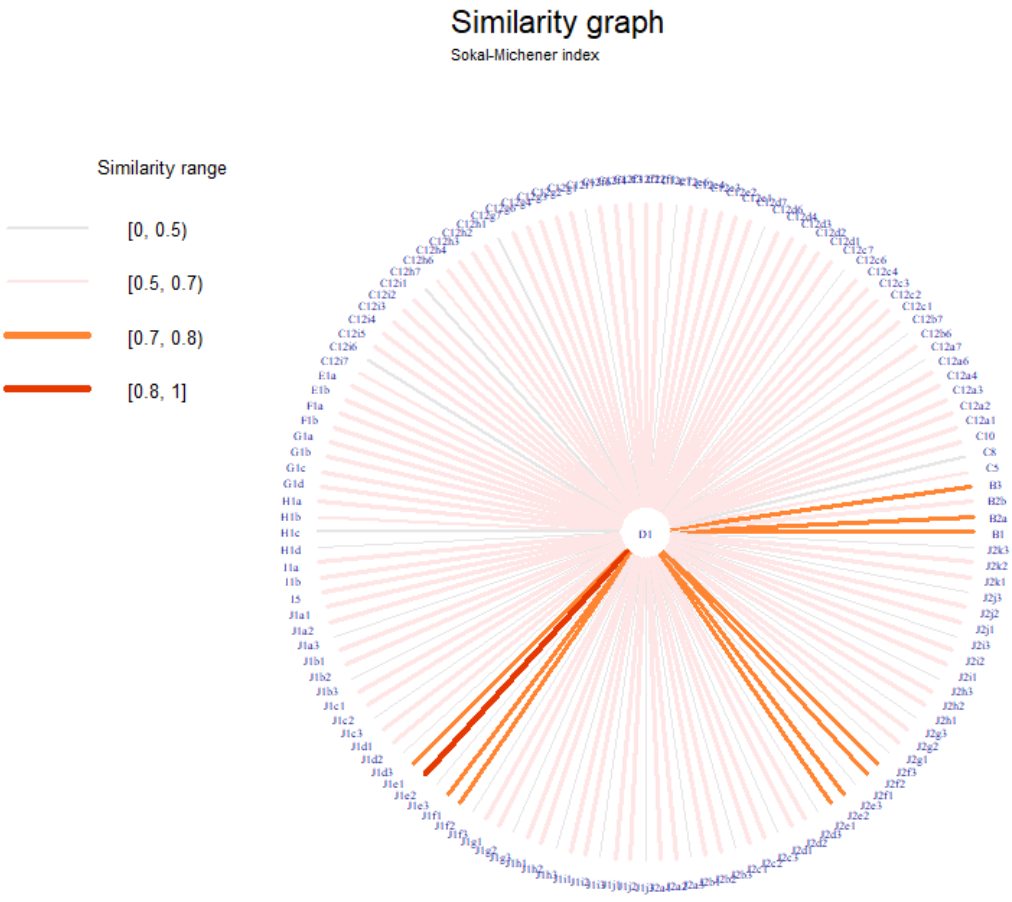

Figure 2: Similarity graph between D1 and all other variables. Edges are coloured depending on the similarity strength.

J1e1-3 and J2e1-3, but we leave those regarding investments in web applications.
Finally, also B1, B2a and B3 are similar to D1: B1 selects the enterprises that employ IT specialists and, among 6918 that do so, 3900 also used cloud computing services; likewise, B2a and B3 describe IT specialists training and hiring and have a great impact on variable D1. We report the three contingency tables of B1, B2a and B3 with D1:
We will keep these variables since they are not expressing the same information of variable D1, although they have a high number of co-presences and co-absences.

At this point, our dataset is 20969×152 long. We look at three barplots to

		D1	
		no	yes
B1	no	10823	3228
	yes	3018	3900

		D1	
		no	yes
B2a	no	12438	4641
	yes	1403	2487

		D1	
		no	yes
B3	no	12833	5094
	yes	1008	2034

have a glance at the main distributions: the target (usage of cloud computing services) in Figure 3; the size of the enterprise as captured by revenues in Figure 4; and the geographical location of the enterprise, in Figure 5.

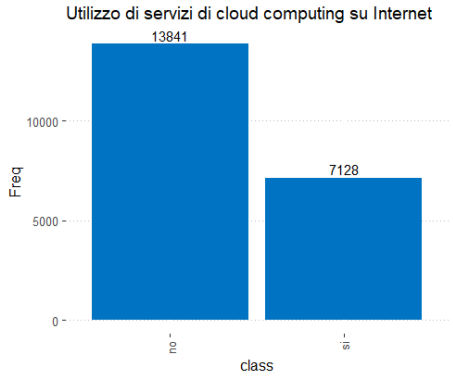

Figure 3: Barplot of `D1` - Usage of Cloud Computing Services

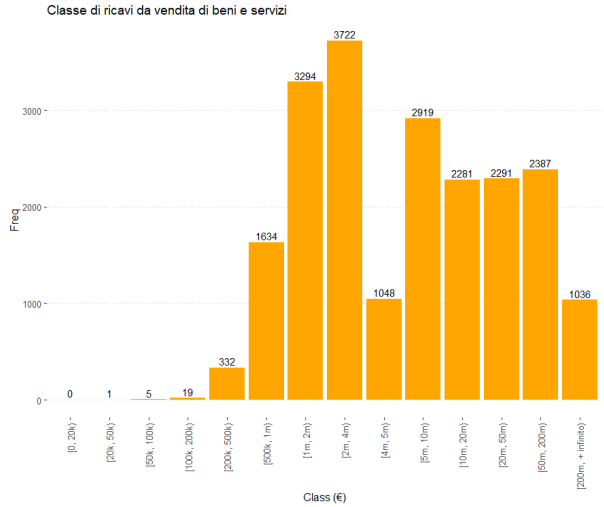

Figure 4: Barplot of `ricavi_cl` - Revenues

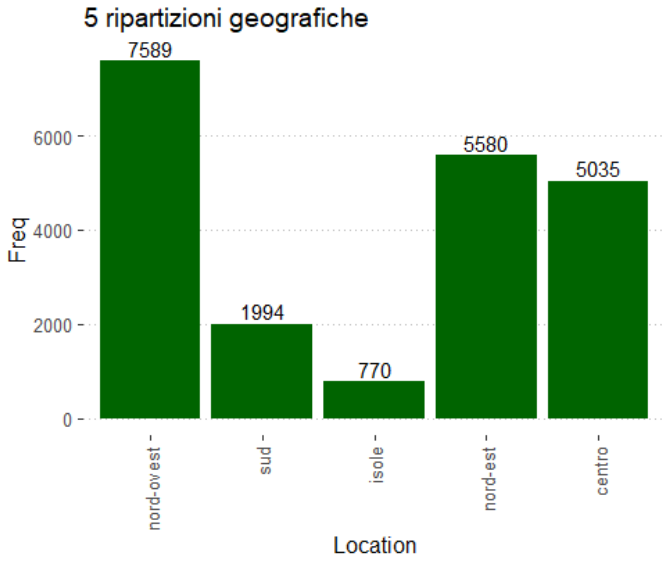

Figure 5: Barplot of `rip` - Geographical Distribution

The target distribution is skewed to "no", with the number of such answers almost doubling the "yes". Few enterprises in this survey has revenues below 500k and finally most respondents are from the north of Italy.

3 Classification

We're now ready to begin the supervised learning task of this project, which is to predict the variable D1, "Usage of cloud computing services". The statistical technique that has been used is logistic regression, which is used in binary classification to predict the class probability. We will give here a brief technical background on this statistical method.

3.1 Logistic regression overview

Let Y_i be a univariate binary (random, in a mathematical framework) variable and let $\underline{X}_i$ be a $d-$dimensional multivariate random vector, such that $\{(X_i, Y_i)\}_{i=1}^n$ is a collection of i.i.d. variables. Suppose that Y_i has a conditional Bernoulli distribution:

$$Y_i|\underline{X}_i \sim B(p(\underline{X}_i))$$

where $p(\cdot)$ is an unknown function of $\underline{X}_i$ that takes values in $[0, 1]$.
Our aim is to estimate this probability function $p(\underline{X})$ and in order to do so we must fix a model. The logistic model makes use of the logistic function and tackles the complexity of the problem by turning it into a parametric one.

In particular, calling $\underline{\beta} = \{\beta_j\}_{j=0}^{d}$ a vector of $d+1$ unknown parameters, the problem is reduced to estimate $\underline{\beta}$ in the following probability function:

$$p(\underline{X}_i, \underline{\beta}) = \frac{e^{\underline{\beta} \cdot (1, \underline{X}_i)}}{1 + e^{\underline{\beta} \cdot (1, \underline{X}_i)}}$$

The estimation is made via MLE. Recalling that a Bernoulli distribution $Y \sim B(p)$ has a likelihood function $\mathcal{L}(y, p) = p^y (1-p)^{1-y}$ and given a sample of i.i.d observations $\{(\underline{x}_i, y_i)\}_{i=1}^{n}$, the optimization problem is:

$$\max_{\underline{\beta}} \mathcal{L}(\underline{x}_i, y_i, \underline{\beta}) = \prod_{i=1}^{n} \left[p(\underline{x}_i, \underline{\beta})\right]^{y_i} \cdot \left[1 - p(\underline{x}_i, \underline{\beta})\right]^{1-y_i}$$

3.2 Logistic regression in practice

In our scenario, the target Y_i is a variable equal to 1 if cloud computing services are used, 0 otherwise (it's 'D1'). The set of numerical variables (cfr. ¶2.1) are injected into the model as they are, contributing numerically to the computation of $p(\underline{x}_i, \underline{\beta})$. The categorical variables are instead turned into dummies, one for each level excluded the baseline. The resulting model has a total of 189 regressors and requires the estimate of 190 coefficients $\{\beta_j\}_{j=0}^{189}$.
After having estimated the function p, classification is obtained by setting a threshold T, in such a way that

$$\hat{y}_i := \begin{cases} 1 \text{ if } p(\underline{x}_i) \geq T \\ 0 \text{ if } p(\underline{x}_i) < T \end{cases}$$

The predicted class $\hat{y}_i$ may differ from the true one y_i and different evaluation metrics are presented in the following paragraph.

Finally, as a good practice in machine learning, the data used for the estimation should be neither used to establish a probability threshold nor to evaluate the goodness of fit. Therefore, the data is split in *training, validation* and *test* sets, of relative size, respectively, 55%, 25% and 20%.

3.3 Evaluation criteria and threshold selection

Several measures or 'scores' are used to assess a classification model's goodness. The most used are certainly *sensitivity*, or true positive rate, that measures the proportion of positives that are correctly identified $\left(\frac{\sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n} y_i}\right)$; *specificity*, or true negative rate, that measure the proportion of negatives that are correctly identified $\left(\frac{\sum_{i=1}^{n} (1-y_i)(1-\hat{y}_i)}{\sum_{i=1}^{n} (1-y_i)}\right)$; and, finally, *accuracy*, that measures the overall proportion of correctly identified observations $\left(\frac{\sum_{i=1}^{n} [y_i \hat{y}_i + (1-y_i)(1-\hat{y}_i)]}{n}\right)$.
A model is better whenever any or all of these metrics are higher. The way we can tune these scores is, besides changing the model, changing the threshold.

Also, a fourth measure will be used in what follows, the Youden Index, defined as:

$$J(T) = \text{sensitivity}(T) + \text{specificity}(T) - 1$$

where T is the threshold used to compute sensitivity and specificity. This index is used for selecting an optimal threshold: the T^* that maximizes it represents a trade-off between optimal sensitivity and optimal specificity and is then regarded as the best choice.
The Youden Index is usually adopted in conjunction with ROC analysis. The ROC curve is a graphic for simultaneously displaying the two types of error for all possible threshold, namely the sensitivity, already introduced, and the *false positive rate*, which is the complimentary of specificity, or $1 -$ specificity. A ROC curve which tends to the upper-left corner is that of a good model, whereas ROC curves close to the bisector signals the low predicting power of the model (a coin-flip model would perform the same). The optimal Youden Index is associated to the point on the ROC curve that lies closest to the top-left corner, as one can verify by rewriting the maximization problem of $J(T)$ as:

$$\max_T J(T) = \text{sensitivity}(T) \; - \; [1 - \text{specificity}(T)] = x - y$$
$$\nabla J = (1, -1)$$

3.4 Results and comment

The MLE estimation on the training set converges in 10 steps, providing a 190-dimensional vector estimate. Each variable's predicting power is tested through a likelihood-ratio test and it turns out that only 34 variables are significant inside the logistic model at the 5% level. In Table 3.4 we report such variables, with the estimated coefficients, standard deviations and value and p-value of likelihood ratio statistics. In the last paragraph we present the scores obtained with an optimal threshold based upon the Youden Index, as described in ¶3.3.

variable	estimate	std.error	statistic	p.value
B1si	0.1852392	0.0738592	2.508003	0.0121416
B2bsi	0.2338341	0.0603467	3.874845	0.0001067
B3si	0.2425271	0.0792103	3.061811	0.0022000
B5dattivita' non svolta	-0.2335726	0.1080224	-2.162261	0.0305981
C2_	0.0056297	0.0009223	6.104276	0.0000000
C4Compresa tra 2Mbit/s e meno di 10Mbit/s	0.5604566	0.2030348	2.760397	0.0057731
C4Compresa tra 10Mbit/s e meno di 30Mbit/s	0.6499985	0.2009393	3.234801	0.0012173
C4Compresa tra 30Mbit/s e meno di 100Mbit/s	0.8532834	0.2019100	4.226059	0.0000238
C4Maggiore o uguale a 100 Mbit/s	0.9075125	0.2065145	4.394425	0.0000111
C5si	0.4571934	0.0714464	6.399110	0.0000000
C8si	0.4052966	0.0819644	4.944788	0.0000008
C10si	0.2670325	0.0575267	4.641891	0.0000035
C12f4si	0.5094104	0.1998990	2.548339	0.0108237
C12g2si	0.3129091	0.1334960	2.343958	0.0190803
E1asi	0.4732862	0.1646191	2.875039	0.0040398
F1bsi	0.2615154	0.1073606	2.435861	0.0148564
G1bsi	0.2898400	0.1212251	2.390924	0.0168060
G1dsi	0.2647514	0.1115559	2.373261	0.0176318
H1bsi	-0.1745743	0.0637216	-2.739642	0.0061506
H1csi	0.3032006	0.0594954	5.096200	0.0000003
I5si	0.2244079	0.0924581	2.427132	0.0152187
J1d1si	-0.3740693	0.1206596	-3.100204	0.0019339
J1f1si	0.3196041	0.0908711	3.517115	0.0004363
J1h1si	-0.4220132	0.1949592	-2.164623	0.0304166
J2f1si	0.5066285	0.1053809	4.807595	0.0000015
J3asi	0.2060746	0.0590139	3.491969	0.0004795
J3csi	0.1620049	0.0734044	2.207018	0.0273128
Ateco_1M	0.4846341	0.1535806	3.155568	0.0016019
Ateco_1N	0.3210695	0.1267376	2.533341	0.0112981
ripsud	-0.2439148	0.0918733	-2.654904	0.0079331
ripcentro	-0.1920110	0.0658445	-2.916129	0.0035440
dom4Appartiene al settore ICT (Ateco 261, 262, 263, 264, 268, 465, 582, 61, 62, 631, 951)	0.3391531	0.1256882	2.698370	0.0069680

Table 1: Significant coefficients of logistic model (5% level)

3.4.1 Estimated coefficients

First, we shall make a comment on remarkable values of some statistically significant coefficients.
We observe an increase in the value of the estimated coefficients through the different levels of variable C4, yielding a higher probability of activating cloud services as the broadband internet connection speed increases.
Variables C5 and C8 show that enterprises that provide mobile devices to their employees and that have their own website have an increased probability of also activating cloud computing services. Interestingly, also the usage of 3D printers (E1a) has a positive impact on D1. Variables G1b and G1d, regarding Big Data usage, contribute positively to D1, indeed one can expect Big Data and cloud computing to be complementary – i.e. used together.
Some counterintuitive results arise from investment-related variables (J***): despite investments in web applications (J1f1, J2f1) seem to play an important role in determining the usage of cloud computing utilities (as anticipated in ¶2.3.2), it looks like investments in other computerized capital goods and in Big Data Analytics in the year 2016 discouraged the usage of cloud computing. Perhaps, enterprises that have invested in these two other assets had their cloud computing investments reduced, and so their usage.
Then, enterprises operating professional, scientific or technical activities (Ateco code 'M'), or offering business support services or belonging to the transport sector (Ateco code 'N') or belonging to the ICT sector (binary variable `dom4`) have more likely activated cloud computing services.
Finally, all dummies for non-northern Italy are negative and two are also significant, signalling that North Italy (the baseline in our variable `rip`), where technological advancement is higher, is also where most enterprises adopt cloud computing services.

The results we've presented here are quite understandable; in particular, it looks like the overall technology advancement of an enterprise, including usage of computers, internet, IT specialists, 3D printing, is the main determinant of the managerial choice to adopt cloud computing services. Sometimes, it might as well be that switching to the cloud is not just a choice, but a need to the enterprise, which has grown (technologically) enough to no longer support on-premise data storage, processing or analysis. On the contrary, it is interesting to observe that many variables did not qualify to bring statistically significant information to the model. One outstanding example of such variables are those related to revenues (`ricavi_cl`): it looks like the usage of cloud computing services is independent on the business size.
The distribution of predicted probabilities in the training set is skewed to the left, as the following figure shows: This skewness reflects the unbalance of the distribution of the true target variable and will affect the threshold choice.

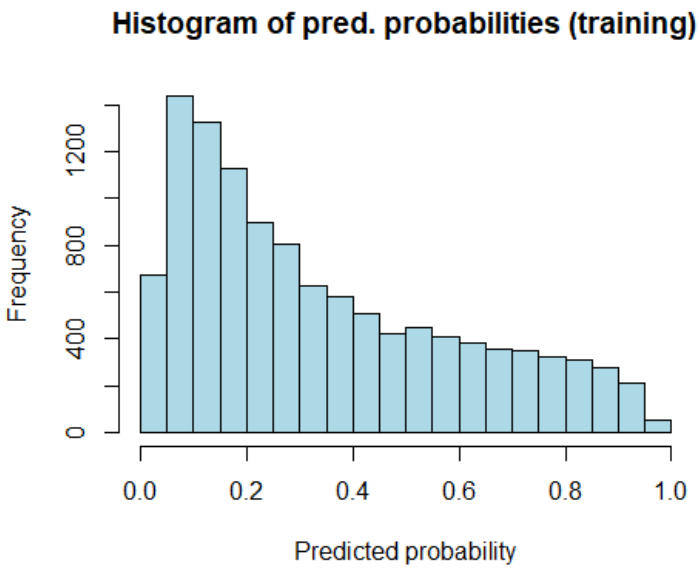

Figure 6: Histogram of the predicted probabilities of the training observations, based on the estimated logistic function.

3.4.2 Classification scores

When it comes to classify observations, a threshold is set according to the maximum Youden index, as discussed previously. In the validation set, across all thresholds T from 0% to 100%, the one that maximises the Youden index is:

$$T^* = 31.68\%$$

The associated maximum Youden index is equal to 0.466. Figure 7 shows the ROC curve and the optimal value for the Youden index.
In the test set, such threshold correctly classifies 3023 out of 4194 observations, achieving an accuracy rate of 72%, with a sensitivity of 72.6% and a specificity of 71.8%. In addition, the proportion of false negative in the predicted negative is $\frac{381}{381+2016} = 15.9\%$, while the proportion of false positive in the predicted positive is $\frac{790}{790+1007} = 44.0\%$: this implies that our model is good for predicting the negative class but less good for predicting the positive class, i.e. given a negative prediction we're more confident that it is correct than if we were given a positive prediction.

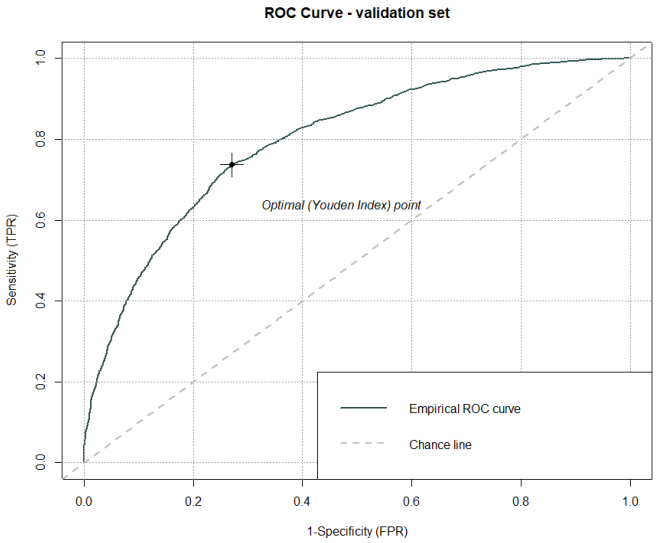

Figure 7: ROC curve based on the validation set

		True	
		no	yes
Predicted	no	2016	381
	yes	790	1007

Table 2: Confusion matrix (test set)

Conclusion

References

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning (with Applications in R)*. Springer, New York.

Appendices

3.5 ICT Survey Variables

VARIABLES OVERVIEW

Code	Type	Levels	Descripion
codice_	Quantitative		Codice progressivo
ricavi_cl	Categorical	14	Classe di ricavi da vendita di beni e servizi
A3_	Quantitative		Percentuale di addetti che usano il computer sul totale degli addetti
B1	Categorical	2	Impiego di addetti specialisti in materie informatiche
B2a	Categorical	2	Corsi di formazione IT destinati agli addetti con competenze specialistiche in ICT
B2b	Categorical	2	Corsi di formazione IT destinati agli addetti senza competenze specialistiche in ICT
B3	Categorical	2	Ha assunto o ha provato ad assumere specialisti ICT
B4	Categorical	2	Difficolta' a coprire i posti vacanti per specialisti ICT
B5a	Categorical	3	Utilizzo prevalente di persone che fanno parte del gruppo di imprese o esterne per Manutenzione delle infrastrutture ICT
B5b	Categorical	3	Utilizzo prevalente di persone che fanno parte del gruppo di imprese o esterne per Supporto per i software di ufficio
B5c	Categorical	3	Utilizzo prevalente di persone che fanno parte del gruppo di imprese o esterne per Sviluppo di sistemi e di software di gestione aziendale
B5d	Categorical	3	Utilizzo prevalente di persone che fanno parte del gruppo di imprese o esterne per Supporto per software e sistemi di gestione aziendale
B5e	Categorical	3	Utilizzo prevalente di persone che fanno parte del gruppo di imprese o esterne per Sviluppo web
B5f	Categorical	3	Utilizzo prevalente di persone che fanno parte del gruppo di imprese o esterne per Supporto per lo sviluppo web
B5g	Categorical	3	Utilizzo prevalente di persone che fanno parte del gruppo di imprese o esterne per Gestione della sicurezza ICT e protezione dei dati
C1	Categorical	2	Connessione ad Internet
C2_	Quantitative		Percentuale di addetti che usano computer connessi Internet sul totale degli addetti
C3	Categorical	2	Tipo di connessione: fissa in banda larga - DSL (xDSL, ADSL, SDSL, VDSL, ecc.), via cavo, fibre ottiche (FTTH), connessioni fisse senza fili, WiFi (anche pubbliche), WiMax
C4	Categorical	5	Velocita' max download connessione fissa in banda larga
C5	Categorical	2	L'impresa fornisce dispositivi portatili con connessione mobile
C6_	Quantitative		Percentuale di addetti provvisti di dispositivi portatili forniti dall'impresa che permettono la connessione mobile ad Internet (in banda larga o meno) sul totale degli addetti
C7a	Categorical	2	connessione mobile per: accedere a sistema di posta elettronica aziendale
C7b	Categorical	2	connessione mobile per: accedere e modificare documenti aziendali
C7c	Categorical	2	connessione mobile per: utilizzare di specifiche applicazioni software aziendali
C8	Categorical	2	Sito web
C9a	Categorical	2	Possibilita' di effettuare ordinazioni o prenotazioni on line (es. carrello della spesa on line)
C9b	Categorical	2	Tracciabilita' on line dell'ordine
C9c	Categorical	2	Accesso a cataloghi di prodotti o listini prezzi
C9d	Categorical	2	Possibilita' di personalizzare i contenuti del sito per i visitatori abituali
C9e	Categorical	2	Possibilita' per i visitatori del sito di personalizzare o progettare prodotti
C9f	Categorical	2	Collegamenti o riferimenti ai profili dell'impresa sui social media
C10	Categorical	2	Pubblicita' a pagamento su Internet
C11a	Categorical	2	Pubblicita a pagamento su Internet, metodi di pubblicita' mirata: ricerca contenuti web
C11b	Categorical	2	Pubblicita a pagamento su Internet, metodi di pubblicita' mirata: tracciabilita' degli utenti
C11c	Categorical	2	Pubblicita a pagamento su Internet, metodi di pubblicita' mirata: geolocalizzazione degli utenti

Code	Type	Levels	Descripion
C11d	Categorical	2	Pubblicita a pagamento su Internet, metodi di pubblicita' mirata: altri metodi
C12a1	Categorical	2	Adempimenti e procedure per il lavoro (INPS/INAIL): Procedure elettroniche troppo complicate e dispendiose in termini di tempo
C12a2	Categorical	2	Adempimenti e procedure per il lavoro (INPS/INAIL): Procedure elettroniche che richiedono ancora la presentazione di documenti cartacei
C12a3	Categorical	2	Adempimenti e procedure per il lavoro (INPS/INAIL): Difficolta' tecniche dipendenti dal sito web/portale (interruzioni procedure, procedure lente)
C12a4	Categorical	2	Adempimenti e procedure per il lavoro (INPS/INAIL): Informazioni insufficienti o poco chiare; mancanza di supporto
C12a5	Categorical	2	Adempimenti e procedure per il lavoro (INPS/INAIL): Timori legati alla sicurezza o alla riservatezza dei dati
C12a6	Categorical	2	Adempimenti e procedure per il lavoro (INPS/INAIL): Nessun problema nell'utilizzo del servizio offerto on-line dalla PA
C12a7	Categorical	2	Adempimenti e procedure per il lavoro (INPS/INAIL): L'impresa non utilizza il servizio e/o si avvale di un intermediario (commercialista, CAF, altra impresa del gruppo, ecc.)
C12b1	Categorical	2	Dichiarazione dei redditi dell'impresa: Procedure elettroniche troppo complicate e dispendiose in termini di tempo
C12b2	Categorical	2	Dichiarazione dei redditi dell'impresa: Procedure elettroniche che richiedono ancora la presentazione di documenti cartacei
C12b3	Categorical	2	Dichiarazione dei redditi dell'impresa: Difficolta' tecniche dipendenti dal sito web/portale (interruzioni procedure, procedure lente)
C12b4	Categorical	2	Dichiarazione dei redditi dell'impresa: Informazioni insufficienti o poco chiare; mancanza di supporto
C12b5	Categorical	2	Dichiarazione dei redditi dell'impresa: Timori legati alla sicurezza o alla riservatezza dei dati
C12b6	Categorical	2	Dichiarazione dei redditi dell'impresa: Nessun problema nell'utilizzo del servizio offerto on-line dalla PA
C12b7	Categorical	2	Dichiarazione dei redditi dell'impresa: L'impresa non utilizza il servizio e/o si avvale di un intermediario (commercialista, CAF, altra impresa del gruppo, ecc.)
C12c1	Categorical	2	Dichiarazione IVA: Procedure elettroniche troppo complicate e dispendiose in termini di tempo
C12c2	Categorical	2	Dichiarazione IVA: Procedure elettroniche che richiedono ancora la presentazione di documenti cartacei
C12c3	Categorical	2	Dichiarazione IVA: Difficolta' tecniche dipendenti dal sito web/portale (interruzioni procedure, procedure lente)
C12c4	Categorical	2	Dichiarazione IVA: Informazioni insufficienti o poco chiare; mancanza di supporto
C12c5	Categorical	2	Dichiarazione IVA: Timori legati alla sicurezza o alla riservatezza dei dati
C12c6	Categorical	2	Dichiarazione IVA: Nessun problema nell'utilizzo del servizio offerto on-line dalla PA
C12c7	Categorical	2	Dichiarazione IVA: L'impresa non utilizza il servizio e/o si avvale di un intermediario (commercialista, CAF, altra impresa del gruppo, ecc.)
C12d1	Categorical	2	Sportello Unico per le attivita' Produttive (permessi di costruire, dichiarazione di inizio attivita', ecc): Procedure elettroniche troppo complicate e dispendiose in termini di tempo
C12d2	Categorical	2	Sportello Unico per le attivita' Produttive (permessi di costruire, dichiarazione di inizio attivita', ecc): Procedure elettroniche che richiedono ancora la presentazione di documenti cartacei
C12d3	Categorical	2	Sportello Unico per le attivita' Produttive (permessi di costruire, dichiarazione di inizio attivita', ecc): Difficolta' tecniche dipendenti dal sito web/portale (interruzioni procedure, procedure len
C12d4	Categorical	2	Sportello Unico per le attivita' Produttive (permessi di costruire, dichiarazione di inizio attivita', ecc): Informazioni insufficienti o poco chiare; mancanza di supporto
C12d5	Categorical	2	Sportello Unico per le attivita' Produttive (permessi di costruire, dichiarazione di inizio attivita', ecc): Timori legati alla sicurezza o alla riservatezza dei dati
C12d6	Categorical	2	Sportello Unico per le attivita' Produttive (permessi di costruire, dichiarazione di inizio attivita', ecc): Nessun problema nell'utilizzo del servizio offerto on-line dalla PA

Code	Type	Levels	Descripion
C12d7	Categorical	2	Sportello Unico per le attivita' Produttive (permessi di costruire, dichiarazione di inizio attivita', ecc): L'impresa non utilizza il servizio e/o si avvale di un intermediario (commercialista, CAF,
C12e1	Categorical	2	Adempimenti e procedure in materia edilizia: Procedure elettroniche troppo complicate e dispendiose in termini di tempo
C12e2	Categorical	2	Adempimenti e procedure in materia edilizia: Procedure elettroniche che richiedono ancora la presentazione di documenti cartacei
C12e3	Categorical	2	Adempimenti e procedure in materia edilizia: Difficolta' tecniche dipendenti dal sito web/portale (interruzioni procedure, procedure lente)
C12e4	Categorical	2	Adempimenti e procedure in materia edilizia: Informazioni insufficienti o poco chiare; mancanza di supporto
C12e5	Categorical	2	Adempimenti e procedure in materia edilizia: Timori legati alla sicurezza o alla riservatezza dei dati
C12e6	Categorical	2	Adempimenti e procedure in materia edilizia: Nessun problema nell'utilizzo del servizio offerto on-line dalla PA
C12e7	Categorical	2	Adempimenti e procedure in materia edilizia: L'impresa non utilizza il servizio e/o si avvale di un intermediario (commercialista, CAF, altra impresa del gruppo, ecc.)
C12f1	Categorical	2	Dichiarazioni doganali (dazi, accise), comunicazioni Intrastat: Procedure elettroniche troppo complicate e dispendiose in termini di tempo
C12f2	Categorical	2	Dichiarazioni doganali (dazi, accise), comunicazioni Intrastat: Procedure elettroniche che richiedono ancora la presentazione di documenti cartacei
C12f3	Categorical	2	Dichiarazioni doganali (dazi, accise), comunicazioni Intrastat: Difficolta' tecniche dipendenti dal sito web/portale (interruzioni procedure, procedure lente)
C12f4	Categorical	2	Dichiarazioni doganali (dazi, accise), comunicazioni Intrastat: Informazioni insufficienti o poco chiare; mancanza di supporto
C12f5	Categorical	2	Dichiarazioni doganali (dazi, accise), comunicazioni Intrastat: Timori legati alla sicurezza o alla riservatezza dei dati
C12f6	Categorical	2	Dichiarazioni doganali (dazi, accise), comunicazioni Intrastat: Nessun problema nell'utilizzo del servizio offerto on-line dalla PA
C12f7	Categorical	2	Dichiarazioni doganali (dazi, accise), comunicazioni Intrastat: L'impresa non utilizza il servizio e/o si avvale di un intermediario (commercialista, CAF, altra impresa del gruppo, ecc.)
C12g1	Categorical	2	Partecipazioni a gare d'appalto e bandi on-line della PA: Procedure elettroniche troppo complicate e dispendiose in termini di tempo
C12g2	Categorical	2	Partecipazioni a gare d'appalto e bandi on-line della PA: Procedure elettroniche che richiedono ancora la presentazione di documenti cartacei
C12g3	Categorical	2	Partecipazioni a gare d'appalto e bandi on-line della PA: Difficolta' tecniche dipendenti dal sito web/portale (interruzioni procedure, procedure lente)
C12g4	Categorical	2	Partecipazioni a gare d'appalto e bandi on-line della PA: Informazioni insufficienti o poco chiare; mancanza di supporto
C12g5	Categorical	2	Partecipazioni a gare d'appalto e bandi on-line della PA: Timori legati alla sicurezza o alla riservatezza dei dati
C12g6	Categorical	2	Partecipazioni a gare d'appalto e bandi on-line della PA: Nessun problema nell'utilizzo del servizio offerto on-line dalla PA
C12g7	Categorical	2	Partecipazioni a gare d'appalto e bandi on-line della PA: L'impresa non utilizza il servizio e/o si avvale di un intermediario (commercialista, CAF, altra impresa del gruppo, ecc.)
C12h1	Categorical	2	Utilizzo della fatturazione elettronica con la PA: Procedure elettroniche troppo complicate e dispendiose in termini di tempo
C12h2	Categorical	2	Utilizzo della fatturazione elettronica con la PA: Procedure elettroniche che richiedono ancora la presentazione di documenti cartacei
C12h3	Categorical	2	Utilizzo della fatturazione elettronica con la PA: Difficolta' tecniche dipendenti dal sito web/portale (interruzioni procedure, procedure lente)
C12h4	Categorical	2	Utilizzo della fatturazione elettronica con la PA: Informazioni insufficienti o poco chiare; mancanza di supporto

Code	Type	Levels	Descripion
C12h5	Categorical	2	Utilizzo della fatturazione elettronica con la PA: Timori legati alla sicurezza o alla riservatezza dei dati
C12h6	Categorical	2	Utilizzo della fatturazione elettronica con la PA: Nessun problema nell'utilizzo del servizio offerto on-line dalla PA
C12h7	Categorical	2	Utilizzo della fatturazione elettronica con la PA: L'impresa non utilizza il servizio e/o si avvale di un intermediario (commercialista, CAF, altra impresa del gruppo, ecc.)
C12i1	Categorical	2	Utilizzo della PEC per interagire con la PA: Procedure elettroniche troppo complicate e dispendiose in termini di tempo
C12i2	Categorical	2	Utilizzo della PEC per interagire con la PA: Procedure elettroniche che richiedono ancora la presentazione di documenti cartacei
C12i3	Categorical	2	Utilizzo della PEC per interagire con la PA: Difficolta' tecniche dipendenti dal sito web/portale (interruzioni procedure, procedure lente)
C12i4	Categorical	2	Utilizzo della PEC per interagire con la PA: Informazioni insufficienti o poco chiare; mancanza di supporto
C12i5	Categorical	2	Utilizzo della PEC per interagire con la PA: Timori legati alla sicurezza o alla riservatezza dei dati
C12i6	Categorical	2	Utilizzo della PEC per interagire con la PA: Nessun problema nell'utilizzo del servizio offerto on-line dalla PA
C12i7	Categorical	2	Utilizzo della PEC per interagire con la PA: L'impresa non utilizza il servizio e/o si avvale di un intermediario (commercialista, CAF, altra impresa del gruppo, ecc.)
D1	Categorical	2	Utilizzo di servizi di cloud computing su Internet
D2a	Categorical	2	Tipologia servizi cloud utilizzati: servizi di posta elettronica
D2b	Categorical	2	Tipologia servizi cloud utilizzati: software per ufficio
D2c	Categorical	2	Tipologia servizi cloud utilizzati: hosting di database dell'impresa
D2d	Categorical	2	Tipologia servizi cloud utilizzati: archiviazione di file
D2e	Categorical	2	Tipologia servizi cloud utilizzati: applicazioni software di finanza e contabilita'
D2f	Categorical	2	Tipologia servizi cloud utilizzati: applicazioni software CRM per gestire le informazioni relative ai propri clienti
D2g	Categorical	2	Tipologia servizi cloud utilizzati: potenza di calcolo per eseguire il software dell'impresa
D3a	Categorical	2	Tipologia di fornitura di servizi cloud: server condivisi (cloud pubblico)
D3b	Categorical	2	Tipologia di fornitura di servizi cloud: server riservati (cloud privato)
E1a	Categorical	2	Utilizzo di stampanti 3D di proprieta', affittate o prese in leasing dall'impresa
E1b	Categorical	2	Utilizzo di servizi di stampa 3D forniti da altre imprese diverse dalla rispondente
E2a	Categorical	2	stampare in 3D prototipi o modelli da vendere
E2b	Categorical	2	stampare in 3D prototipi o modelli per uso interno all'impresa
E2c	Categorical	2	stampare in 3D beni da vendere esclusi prototipi o modelli
E2d	Categorical	2	stampare in 3D beni da utilizzare nel processo di produzione della vostra impresa escluso prototipi o modelli
F1a	Categorical	2	Robot industriali
F1b	Categorical	2	Robot di servizio
F2a	Categorical	2	Robot di servizio: compiti di sorveglianza, sicurezza o ispezione
F2b	Categorical	2	Robot di servizio: trasporto di persone o beni
F2c	Categorical	2	Robot di servizio: attivita' di pulizia o di smaltimento dei rifiuti
F2d	Categorical	2	Robot di servizio: sistemi di gestione del magazzino
F2e	Categorical	2	Robot di servizio: lavori di assemblaggio
F2f	Categorical	2	Robot di servizio: compiti da impiegato di un negozio robotizzato
F2g	Categorical	2	Robot di servizio: lavori di costruzione o riparazione di danni
G1a	Categorical	2	Big Data Analisys - fonti di dati: dispositivi intelligenti o sensori

Code	Type	Levels	Descripion
G1b	Categorical	2	Big Data Analisys - fonti di dati: geolocalizzazione derivante dall'utilizzo di dispositivi portatili
G1c	Categorical	2	Big Data Analisys - fonti di dati: social media
G1d	Categorical	2	Big Data Analisys - fonti di dati: altro
G2a	Categorical	2	Big Data Analisys: personale interno
G2b	Categorical	2	Big Data Analisys: personale esterno
H1a	Categorical	2	Fatture elettroniche adatte B2B e B2C
H1b	Categorical	2	Fatture elettroniche adatte B2G
H1c	Categorical	2	Fatture in formato elettronico non adatte
H1d	Categorical	2	Fatture cartacee
H2	Categorical	5	Classe percentuale delle fatture elettroniche inviate in formato standard
I1a	Categorical	2	Vendite web: tramite siti web o app dell'impresa
I1b	Categorical	2	Vendite web: tramite siti web di intermediari ovvero siti di ecommerce (marketplace) o app
I2_cl	Categorical	7	Classe percentuale degli ordini di vendita sito web rispetto al totale dei ricavi
I3a	Quantitative		Percentuale delle vendite web destinate a consumatori privati (B2C)
I3b	Quantitative		Percentuale delle vendite web destinate ad altre imprese (B2B) o alla Pubblica Amministrazione (B2G)
I4a	Quantitative		Vendite web: tramite siti web o app dell'impresa
I4b	Quantitative		Vendite web: tramite siti web di intermediari ovvero siti di ecommerce (marketplace) o app
I5	Categorical	2	Ordini di vendita attraverso sistemi di tipo EDI
I6_cl	Categorical	7	Classe percentuale degli ordini di vendita via EDI rispetto al totale dei ricavi
J1a1	Categorical	2	Investimenti 2016: Soluzioni di Internet delle cose o IoT (ad es. Rfid, sensori, oggetti conne
J1a2	Categorical	2	Investimenti 2017: Soluzioni di Internet delle cose o IoT (ad es. Rfid, sensori, oggetti conne
J1a3	Categorical	2	Nessun investimento 2016-2017: Soluzioni di Internet delle cose o IoT (ad es. Rfid, sensori, oggetti conne
J1b1	Categorical	2	Investimenti 2016: Stampa 3D
J1b2	Categorical	2	Investimenti 2017: Stampa 3D
J1b3	Categorical	2	Nessun investimento 2016-2017: Stampa 3D
J1c1	Categorical	2	Investimenti 2016: Robotica (robot industriali, robot collaborativi interconnessi e programmabili)
J1c2	Categorical	2	Investimenti 2017: Robotica (robot industriali, robot collaborativi interconnessi e programmabili)
J1c3	Categorical	2	Nessun investimento 2016-2017: Robotica (robot industriali, robot collaborativi interconnessi e programmabili)
J1d1	Categorical	2	Investimenti 2016: Altri beni strumentali computerizzati o gestiti tramite sensori e interconnessi con altri sistemi aziendali
J1d2	Categorical	2	Investimenti 2017: Altri beni strumentali computerizzati o gestiti tramite sensori e interconnessi con altri sistemi aziendali
J1d3	Categorical	2	Nessun investimento 2016-2017: Altri beni strumentali computerizzati o gestiti tramite sensori e interconnessi con altri sistemi aziendali
J1e1	Categorical	2	Investimenti 2016: Cloud Computing (insieme di servizi informatici utilizzabili tramite Internet che consentono l'accesso a software, potenza di calcolo, capacita' di memoria, ecc.)
J1e2	Categorical	2	Investimenti 2017: Cloud Computing (insieme di servizi informatici utilizzabili tramite Internet che consentono l'accesso a software, potenza di calcolo, capacita' di memoria, ecc.)
J1e3	Categorical	2	Nessun investimento 2016-2017: Cloud Computing (insieme di servizi informatici utilizzabili tramite Internet che consentono l'accesso a software, potenza di calcolo, capacita' di memoria, ecc.)
J1f1	Categorical	2	Investimenti 2016: Applicazioni web o app (applicazioni accessibili via Internet comprese quelle gestionali)
J1f2	Categorical	2	Investimenti 2017: Applicazioni web o app (applicazioni accessibili via Internet comprese quelle gestionali)

Code	Type	Levels	Descripion
J1f3	Categorical	2	Nessun investimento 2016-2017: Applicazioni web o app (applicazioni accessibili via Internet comprese quelle gestionali)
J1g1	Categorical	2	Investimenti 2016: Vendite online
J1g2	Categorical	2	Investimenti 2017: Vendite online
J1g3	Categorical	2	Nessun investimento 2016-2017: Vendite online
J1h1	Categorical	2	Investimenti 2016: Big Data Analytics (uso di tecniche, tecnologie e software per l'analisi di grandi quantita' di dati)
J1h2	Categorical	2	Investimenti 2017: Big Data Analytics (uso di tecniche, tecnologie e software per l'analisi di grandi quantita' di dati)
J1h3	Categorical	2	Nessun investimento 2016-2017: Big Data Analytics (uso di tecniche, tecnologie e software per l'analisi di grandi quantita' di dati)
J1i1	Categorical	2	Investimenti 2016: Realta' aumentata e realta' virtuale
J1i2	Categorical	2	Investimenti 2017: Realta' aumentata e realta' virtuale
J1i3	Categorical	2	Nessun investimento 2016-2017: Realta' aumentata e realta' virtuale
J1j1	Categorical	2	Investimenti 2016: Sicurezza informatica
J1j2	Categorical	2	Investimenti 2017: Sicurezza informatica
J1j3	Categorical	2	Nessun investimento 2016-2017: Sicurezza informatica
J2a1	Categorical	2	Investimenti 2018: Soluzioni di Internet delle cose o IoT (ad es. Rfid, sensori, oggetti conne
J2a2	Categorical	2	Investimenti 2019: Soluzioni di Internet delle cose o IoT (ad es. Rfid, sensori, oggetti conne
J2a3	Categorical	2	Nessun investimento 2018-2019: Soluzioni di Internet delle cose o IoT (ad es. Rfid, sensori, oggetti conne
J2b1	Categorical	2	Investimenti 2018: Stampa 3D
J2b2	Categorical	2	Investimenti 2019: Stampa 3D
J2b3	Categorical	2	Nessun investimento 2018-2019: Stampa 3D
J2c1	Categorical	2	Investimenti 2018: Robotica (robot industriali, robot collaborativi interconnessi e programmabili)
J2c2	Categorical	2	Investimenti 2019: Robotica (robot industriali, robot collaborativi interconnessi e programmabili)
J2c3	Categorical	2	Nessun investimento 2018-2019: Robotica (robot industriali, robot collaborativi interconnessi e programmabili)
J2d1	Categorical	2	Investimenti 2018: Altri beni strumentali computerizzati o gestiti tramite sensori e interconnessi con altri sistemi aziendali
J2d2	Categorical	2	Investimenti 2019: Altri beni strumentali computerizzati o gestiti tramite sensori e interconnessi con altri sistemi aziendali
J2d3	Categorical	2	Nessun investimento 2018-2019: Altri beni strumentali computerizzati o gestiti tramite sensori e interconnessi con altri sistemi aziendali
J2e1	Categorical	2	Investimenti 2018: Cloud Computing (insieme di servizi informatici utilizzabili tramite Internet che consentono l'accesso a software, potenza di calcolo, capacita' di memoria, ecc.)
J2e2	Categorical	2	Investimenti 2019: Cloud Computing (insieme di servizi informatici utilizzabili tramite Internet che consentono l'accesso a software, potenza di calcolo, capacita' di memoria, ecc.)
J2e3	Categorical	2	Nessun investimento 2018-2019: Cloud Computing (insieme di servizi informatici utilizzabili tramite Internet che consentono l'accesso a software, potenza di calcolo, capacita' di memoria, ecc.)
J2f1	Categorical	2	Investimenti 2018: Applicazioni web o app (applicazioni accessibili via Internet comprese quelle gestionali)
J2f2	Categorical	2	Investimenti 2019: Applicazioni web o app (applicazioni accessibili via Internet comprese quelle gestionali)
J2f3	Categorical	2	Nessun investimento 2018-2019: Applicazioni web o app (applicazioni accessibili via Internet comprese quelle gestionali)
J2g1	Categorical	2	Investimenti 2018: Vendite online
J2g2	Categorical	2	Investimenti 2019: Vendite online

Code	Type	Levels	Descripion
J2g3	Categorical	2	Nessun investimento 2018-2019: Vendite online
J2h1	Categorical	2	Investimenti 2018: Big Data Analytics (uso di tecniche, tecnologie e software per l'analisi di grandi quantita' di dati)
J2h2	Categorical	2	Investimenti 2019: Big Data Analytics (uso di tecniche, tecnologie e software per l'analisi di grandi quantita' di dati)
J2h3	Categorical	2	Nessun investimento 2018-2019: Big Data Analytics (uso di tecniche, tecnologie e software per l'analisi di grandi quantita' di dati)
J2i1	Categorical	2	Investimenti 2018: Realta' aumentata e realta' virtuale
J2i2	Categorical	2	Investimenti 2019: Realta' aumentata e realta' virtuale
J2i3	Categorical	2	Nessun investimento 2018-2019: Realta' aumentata e realta' virtuale
J2j1	Categorical	2	Investimenti 2018: Sicurezza informatica
J2j2	Categorical	2	Investimenti 2019: Sicurezza informatica
J2j3	Categorical	2	Nessun investimento 2018-2019: Sicurezza informatica
J2k1	Categorical	2	Investimenti 2018: Altro
J2k2	Categorical	2	Investimenti 2019: Altro
J2k3	Categorical	2	Nessun investimento 2018-2019: Altro
J3a	Categorical	2	Fattori digitalizzazione 2018-2019: Infrastruttura e connessione in banda ultralarga
J3b	Categorical	2	Fattori digitalizzazione 2018-2019: Agevolazioni, finanziamenti, incentivi fiscali a sostegno della digitalizzazione
J3c	Categorical	2	Fattori digitalizzazione 2018-2019: Iniziative digitali della pubblica amministrazione
J3d	Categorical	2	Fattori digitalizzazione 2018-2019: Capacita' di fare rete attuando modelli di collaborazione con altre imprese e centri di ricerca per la digitalizzazio
J3e	Categorical	2	Fattori digitalizzazione 2018-2019: Inserimento/sviluppo di nuove competenze digitali
J3f	Categorical	2	Fattori digitalizzazione 2018-2019:Sviluppo/consolidamento di competenze di personale gia' esistente
J3g	Categorical	2	Fattori digitalizzazione 2018-2019: Sviluppo di una strategia di digitalizzazione dell'impresa
J3h	Categorical	2	Fattori digitalizzazione 2018-2019: Altro
J3i	Categorical	2	Fattori digitalizzazione 2018-2019: Nessun fattore di digitalizzazione puo' incidere
J3j	Categorical	2	Fattori digitalizzazione 2018-2019: Non so
Ateco_1	Categorical	12	Ateco, classificazione Ateco2007 a una lettera (contiene missing)
mac	Categorical	4	Macrosettore
clad3	Categorical	3	Classe addetti
rip	Categorical	5	5 ripartizioni geografiche
dom4	Categorical	3	Dominio relativo al settore ICT che comprende le seguenti attivita' economiche: 261+262+263+264+268+465+582+61+62+631+951
coeffin			Peso da utilizzare per riporto all'universo
I2I6_cl	Categorical	7	Classe percentuale degli ordini di vendita online

3.6 R Code

The following R code has been used to produce the results in this paper. In order to replicate such results, please note that:

- The data attached to this paper, named *ICT_Microdati_2018.txt*, is to be put in the same directory of the R code.
- The acquisition, factorization and labelling is done in file *PGM_2018_IT_DELIMITED.R*, which is not attached in this appendix because of its length. The dataframe *DF_ICT_A2018* is produced by the aforesaid R script.
- The code has been divided in 2 files for simplicity, besides acquisition. Follow the same order of execution provided here.

```r
#########################  EXPLORATORY ANALYSIS  ##############################
library(ggplot2)
library(dplyr)
library(ggpubr)
library(stringr)
library(corrplot)
library(proxy)
library(igraph)
library(jtools)
library(kableExtra)

theme_set(theme_pubr())

df <- DF_ICT_A2018

### === FEATURE STUDY === ###
str(df[,1:25])
str(df[,26:50])
str(df[,51:75])
str(df[,76:100])
str(df[,101:125])
str(df[,126:150])
str(df[,151:175])
str(df[,176:200])
str(df[,201:ncol(df)])

update_lab_type <- function(){
  lab_type <<- rep("", ncol(df))
  lab_levels <<- rep("", ncol(df))
  bool.factor <<- sapply(df, is.factor)
  bool.numeric <<- !bool.factor & sapply(df, is.numeric)
  bool.binary.yn <<- sapply(df, function(x) (is.factor(x) && identical(levels(x), c("no",
  "si"))))

  lab_type[bool.factor] <<- "Categorical"
  lab_type[bool.numeric] <<- "Quantitative"
  lab_levels[bool.factor] <<- sapply(df[,bool.factor], function(x){ length(levels(x))})

  which.factor <<- names(bool.factor[bool.factor == TRUE])
  which.numeric <<- names(bool.numeric[bool.numeric == TRUE])
  which.binary.yn <<- names(bool.binary.yn[bool.binary.yn == TRUE])
}

update_lab_type()

# Create variable table
varlab_df <- data.frame("Code" =names(varlab), "Type" =lab_type,
                        "Levels" = lab_levels, "Descripion" = unname(varlab))
varlab_df %>%
  kbl() %>%
  kable_styling(bootstrap_options = c("striped", "condensed"))


## == DROP IDENTIFIERS AND IRRELEVANT VARIABLES == ##
df <- df %>%
  select(!c("codice_", "coeffin"))


## == Numeric features == ##
varlab[which.numeric]
cor_numeric <- cor(df[,which.numeric], use="pairwise.complete.obs")
corrplot(cor_numeric)
# I3a and I3b are perfectly negatively correlated! Let's drop I3b
# I4a and I4b are perfectly negatively correlated! Let's drop I4b
df <- df %>%
  select(!c("I3b", "I4b", "A3_"))
update_lab_type()

cor_numeric <- cor(df[,which.numeric], use="pairwise.complete.obs")
```

```r

## == NAs study == ##
na_threshold <- 0.05
sumna <- function(x){ sum(is.na(x)) }
percna <- function(x) {
  if(is.vector(x) || is.factor(x)) {
    return(sumna(x)/length(x))
  } else if(is.data.frame(x) || is.matrix(x)) {
    return(sumna(x)/prod(dim(x)))
  }
}

# = Column NAs = #
colna <- apply(df, 2, sumna) # NAs in each column
perccolna <- apply(df, 2, percna) # percentage of NAs in each column
highNAcol <- perccolna[perccolna >= na_threshold]
print(names(highNAcol))
varlab[names(highNAcol)]
#View(data.frame("valore" = highNAcol, "label"=unname(varlab[names(highNAcol)])))
df <- df[, perccolna < na_threshold]

# = Row NAs = #
rowna <- apply(df, 1, sumna) # NAs in each row
percrowna <- apply(df, 1, percna) # percentage of NAs in each row
sum(percrowna > 0)
#df <- df[percrowna < na_threshold,]
df <- df[complete.cases(df),]


### === CATEGORICAL FEATURES === #

## == Constant columns == ##
# Which columns have constant values?
names(which(apply(df, 2, function(x) all(duplicated(x[!is.na(x)])[-1L]))))
df <- df %>%
  select(!c("C1", "C3"))

## == Similarity analysis == ##
get_similar_cols <- function(threshold=0.99) {
  update_lab_type()
  n <- length(which.binary.yn)
  out1 <- matrix(NA, nrow=n, ncol=n, dimnames=list(which.binary.yn, which.binary.yn))
  out2 <- list()
  ii <- 1
  for(i in seq(1,length(which.binary.yn)-1)) {
    for(j in seq(i, length(which.binary.yn))) {
      if( all(levels(df[,which.binary.yn[i]]) == levels(df[,which.binary.yn[j]])) ) {
        # Sokal-Michener index (simple matching)
        out1[i,j] <- out1[j,i] <- sum(df[,which.binary.yn[i]] == df[,which.binary.yn[j]]) /
        nrow(df)
        if((i != j) && out1[i,j] > threshold) {
          out2[[ii]] <- c(which.binary.yn[i], which.binary.yn[j])
          ii <- ii+1
        }
      }
    }
  }
  return(list("similarity_matrix"=out1, "similar_cols"=out2))
}

#similar_cols <- get_similar_cols(0.5)
similar_cols <- get_similar_cols(0.99)
similar_cols$similar_cols

# The following columns are pairwise similar (>0.99).
# Drop the repetitions:
pairwise_similar_drop <- c("C12a5", "C12b1", "C12b2", "C12b3", "C12b4", "C12b5", "C12c5",
"C12d5", "C12e5", "C12f5", "C12g5", "C12h5")
```

```r
df <- df %>%
  select(-pairwise_similar_drop)

similar_cols <- get_similar_cols()
similar_cols$similar_cols

# Similarity with the target
update_lab_type()
D1_similar <- matrix(0, nrow=length(which.binary.yn), ncol=length(which.binary.yn),
dimnames=list(which.binary.yn, which.binary.yn))
D1_similar["D1",] <- similar_cols$similarity_matrix["D1",]
D1_similar[,"D1"] <- similar_cols$similarity_matrix[,"D1"]
D1_similar <- na.omit(D1_similar)

net <- graph.adjacency(D1_similar,mode="undirected",weighted=TRUE,diag=FALSE)
E(net)$edge.color <- "#e63900"
E(net)$edge.color[E(net)$weight < 0.8] <- "#ff8533"
E(net)$edge.color[E(net)$weight < 0.7] <- "#ffe6e6"
E(net)$edge.color[E(net)$weight < 0.5] <- "#e6e6e6"
E(net)$width <- E(net)$weight*5
V(net)$label.cex=0.5
plot.igraph(net, vertex.label=V(net)$name, layout=layout_as_star(net,center = "D1"),
            vertex.shape="none", edge.color=E(net)$edge.color)
mtext(side=3, line=3, at=-0.55, adj=0, cex=1.4, "Similarity graph")
mtext(side=3, line=2, at=-0.55, adj=0, cex=0.7, "Sokal-Michener index")
legend("topleft", c("[0, 0.5)","[0.5, 0.7)", "[0.7, 0.8)", "[0.8, 1]"), bty = "n",
       lwd = c(2,2,5,5), cex = 0.8, col = c("#e6e6e6", "#ffe6e6", "#ff8533", "#e63900"),
       lty = 1, pch = NA, title="Similarity range")

table(df$B1, df$D1, dnn=list("B1", "D1"))
table(df$B2a, df$D1, dnn=list("B1", "D1"))
table(df$B3, df$D1, dnn=list("B1", "D1"))


# Drop selected columns similar to the target (cloud computing investments)
target_similar_drop <- c("J1e1", "J1e2", "J1e3", "J2e1", "J2e2", "J2e3")
# View(data.frame(varlab[target_similar_drop]))
df <- df %>%
  select(-target_similar_drop)
update_lab_type()

### === Target barplot === ###
D1_tab <- table(df$D1, dnn=c("class")) # for NAs insert argument: useNA = "always"
D1_df <- data.frame(D1_tab)

ggplot(D1_df, aes(x=class, y=Freq)) +
  geom_bar(fill = "#0073C2FF", stat = "identity") +
  geom_text(aes(label = Freq), vjust = -0.3) +
  theme_pubclean() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(title = varlab["D1"])

### === Revenues Histogram === ###
ricavi_tab <- table(df$ricavi_cl, dnn=c("class")) # for NAs insert argument: useNA = "always"
nastring <- str_wrap("missing per esigenze di anonimizzazione", width=25)
ricavi_df <- data.frame(ricavi_tab)

ggplot(ricavi_df, aes(x=class, y=Freq)) +
  geom_bar(fill = "orange", stat = "identity") +
  geom_text(aes(label = Freq), vjust = -0.3) +
  theme_pubclean() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(title = varlab["ricavi_cl"]) +
  xlab("Class (€)")

# Just one element for ricavi_cl class "[20k, 50k)": drop it
df <- df %>%
  filter(ricavi_cl != "[20k, 50k)")
df$ricavi_cl <- droplevels(df$ricavi_cl)
```

```r
update_lab_type()


### Geographic histrogram ###
rip_tab <- table(df$rip, dnn=c("class")) # for NAs insert argument: useNA = "always"
rip_df <- data.frame(rip_tab)

ggplot(rip_df, aes(x=class, y=Freq)) +
  geom_bar(fill = "darkgreen", stat = "identity") +
  geom_text(aes(label = Freq), vjust = -0.3) +
  theme_pubclean() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(title = varlab["rip"]) +
  xlab("Location")

```

```r
##########################  CLASSIFICATION  ##############################
library(ROCit)
library(ggstance)
library(kableExtra)

### === TRAINING/VALIDATION/TEST SET SPLIT === ###
set.seed(130497)

train_perc <- 0.55
val_perc <- 0.25
test_perc <- 0.20
train_size <- floor(train_perc * nrow(df))
val_size <- floor(val_perc * nrow(df))
test_size <- floor(test_perc * nrow(df))
train <- sort(sample(seq_len(nrow(df)), size=train_size))
ntrain <- setdiff(seq_len(nrow(df)), train)
val <- sort(sample(ntrain, size=val_size))
test <- setdiff(ntrain, val)



### === LOGISTIC MODEL === ###
logm <- glm(formula = D1 ~ ., data=df[train,], family=binomial())
probs <- predict(logm, newdata=df, type="response")
true <- df$D1

scores <- list()
i <- 1
for(set in list(train, val, test)){
  scores[[i]] <- measureit(score=probs[set],
                           class=true[set],
                           measure=c("ACC", "SENS", "SPEC"))
  # Youden Index
  scores[[i]]$YOUD <- scores[[i]]$SPEC + scores[[i]]$SENS - 1
  i <- i+1
}

# 1 = train
# 2 = val
# 3 = test

## == ESTIMATED COEFFICIENTS == ##
dfm <- generics::tidy(logm)
colnames(dfm)[1] <- "variable"

conf_level <- 0.05

dfm_signif <- dfm[dfm$p.value < conf_level,]
print(dfm_signif, n=1e5)
dfm_signif %>%
  kbl(format="latex") %>%
  kable_classic_2(full_width = F) %>%
  column_spec(2, color = ifelse(dfm_signif$estimate > 0, "#005ce6", "#cc2900"))

dfm_n_signif <- dfm[dfm$p.value >= conf_level, ]
print(dfm_n_signif, n=1e5)

hist(probs[train], main = "Histogram of pred. probabilities (training)",
     xlab="Predicted probability", col="lightblue")

#plot_coefs(dfm_signif)


## == HYPERPARAMETER TUNING (Threshold Selection) == ##

# Optimal Youden Index
OYI_val <- max(scores[[2]]$YOUD)
OYIpos_val <- which.max(scores[[2]]$YOUD)
OYIthreshold <- scores[[2]]$Cutoff[OYIpos_val]
```

```r
print(c("OYI"=OYI_val, "Optimal threshold"=OYIthreshold))
# = ROC curve = #
plot(rocit(score=probs[val], class=true[val], negref="no"))
title("ROC Curve - validation set")

preds <- rep("no", length(probs))
preds[probs > OYIthreshold] <- "si"

## == SCORES == ##
# = Test set = #
pos_test <- which.min(abs(scores[[3]]$Cutoff-OYIthreshold))
# ACCURACY
scores[[3]]$ACC[pos_test]
# SENSITIVITY
scores[[3]]$SENS[pos_test]
# SPECIFICITY
scores[[3]]$SPEC[pos_test]

table(na.omit(df[test,"D1"]), preds[test],  dnn=c("true", "predicted"))
View(data.frame(probs[test], true[test]))
```