Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230682

Hipertension Demand Forecasting Using Cross-Correlation and Lagged Multiple Linear Regression Models for Anticipating Health Resources Needs

Guillem HERNÁNDEZ GUILLAMET [a,b,c,1], Beatriz LÓPEZ [c] and Oriol ESTRADA [a,b] and Francesc LÓPEZ SEGUÍ [a,b]

[a] *Research Group on Innovation, Health Economics and Digital Transformation (Institut de Recerca Germans Trias i Pujol)*
[b] *Hospital Universitari Germans Trias i Pujol (Institut Català de la Salut)*
[c] *eXiT research group, University of Girona*
ORCiD ID: Guillem Hernández Guillamet https://orcid.org/0000-0002-1053-3084, Beatriz López https://orcid.org/0000-0001-9210-0073, Oriol Estrada https://orcid.org/0000-0003-1414-5000, Francesc López Seguí https://orcid.org/0000-0003-0977-0215

Abstract. This article presents an algorithm that uses a combination of cross-correlation analysis and lagged multiple linear regression models to predict the time-series of future demand for clinical visits associated with a certain diagnosis, specifically hypertension, in the Catalan health-care system. The algorithm aims to provide a robust and explainable feature selection set of predictors. The study demonstrates that it is possible to predict demand associated with a diagnosis through the demand for previous clinical visits, and identifies important predictors for example case hypertension-related visits. The data used is from the primary care services of the Catalan Institute of Health, and the methodology can be applied to optimize resource allocation in the healthcare system.

Keywords. multiple linear regression, forecasting, healthcare demand

1. Introduction

Many studies and algorithms have explored the possibility of predicting future demand to anticipate resource consumption and optimize their distribution. The case study is particularly important in the healthcare sector, which is heavily strained by the shortage of professionals and the aging population in many developed countries. In the case of the Catalan healthcare system, this problem is exacerbated due to systemic underfunding, overload and talent drain of professionals, entrenched waiting lists, the commercializa-

[1]Corresponding Author: Guillem Hernández Guillamet (ghernandezgu.germanstrias@gencat.cat ; +34 648977302)

tion of healthcare, and a gradual aging of the population that is living longer (with all its clinical consequences). It is also important to point out the effect that the COVID-19 pandemic has had on the healthcare system. Most of these issues have been aggravated, and the normal functioning of the system has been distorted. With the lockdown, the treatment of multiple conditions and patients has been postponed, and it is necessary to recover the lost activity [1][2]. In this context, the ability to predict the future demand of the healthcare system is essential for the efficient planning of resources and modeling the post-pandemic demand. It is particularly interesting not only to predict the demand but also the reason for consultation, in this case, the diagnoses associated with the medical visit, to distribute healthcare professionals according to their expertise and develop a more flexible system.

In this article, we present an algorithm that uses a combination of cross-correlation analysis and a lagged multiple linear regression (lagged MLR) framework to predict the time series of future demand for clinical visits associated with a certain diagnosis, in this case, hypertension. The algorithm seeks to achieve the best prediction while penalizing the progressive entry of new predictors, trying to achieve the best forecasting with the minimum number of variables. Therefore, the algorithm not only achieves good predictive results but also indicates which diagnoses may be important for predicting future demand for visits associated with hypertension.

This paper is organized as follows. In Section 2 we provide a description of related proposals. Section 3 is devoted to outlining the methodology and the algorithm. In Section 4 the main results of its application are presented and discussed. Finally, section 5 presents the main conclusions of the study.

2. Related work

The hypothesis we aim to demonstrate is that the demand associated with a visit related to a specific diagnosis can be predicted through the demand for previous clinical visits. There are very few studies that focus on developing demand prediction models in the medical field. Most research revolves around predicting epidemiological diagnoses and identifying relevant predictor variables. Some studies attempt to predict epidemiological visits, such as flu outbreaks [3]. Other studies explore the possibility of predicting the spread of COVID-19 using climate variables [4][5][6][7][8]. Additionally, there are studies specifically dedicated to predicting visit demands using Deep Learning [9] or predicting unplanned visits for diabetic patients [10]. Our approach differs from the previous ones since it aims to predict demand, not only cases. Moreover, the use of regression models provides a familiar framework for health professionals and allows for a feature selection procedure that can be later used for prediction with AI methods.

On the other hand, one interesting research related work is the prediction of emergency attendances, since they require from an accurate resource planning as we are trying to pursue. In [11] a combination of MLR and Artifical Neural Networks (ANN) is used to capture the complexity of emergency attendances in a touristic area, with a huge variability between different periods of the year (seasonality). It is evident that seasonality plays a key role, as well as other variables in prediction, but we have not included them due to this ”evidence.” Our work aims to understand which clinical variables act as a proxy for the target, independently of the season.

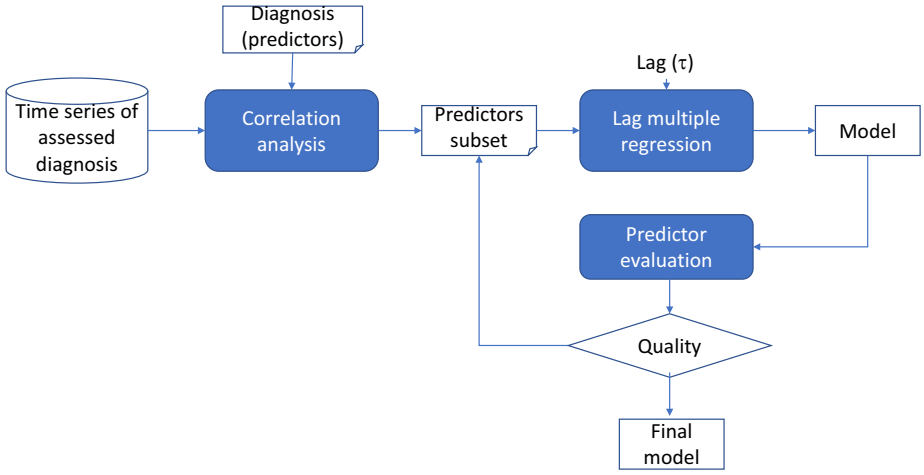

Figure 1. Flow chart of the methodology.

There are some other works that link emmergency departments to some secfic resources, as for example [12]. Linear regression analysis is also used in [12], as we do in our work, showing how simple tools can provide a big improvement in healthcare systems for resource management.

3. Methodology

The framework used can be divided into three major steps (Figure 1). The starting point is time series of diagnoses. Diagnoses play both the role of predictors and target. Our aim is to predict the number of visits associated with a certain diagnosis based on the information of other diagnoses. This prediction is an estimation of the demand for services. Firstly, a cross-correlation analysis is developed between the different time series of demand for diagnoses, and the number of predictors is reduced to a non-correlated subset. Subsequently, a MLR model is trained to predict the target diagnosis. MLR models offer a high degree of interpretability and are easily understandable for healthcare professionals who use them regularly. These models allow for the identification and quantification of the relationship between predictor variables and the target variable, providing a clear understanding of the factors that influence the outcomes. Additionally, their simplicity and familiarity in the healthcare field facilitate their utilization and acceptance among professionals. For these reasons, we have selected these models over other more complex or less interpretable options. This method is applied iteratively; so the MLR model starts with a single predictor and analyzes the progressive entry of new ones. Moreover, the regression model is parameterized by a lag value to build predictive models according to a given time horizon.

With this proposal, we aim to develop a robust model with greater explainability than alternative models such as deep learning (DL). Moreover, we find the minimum number of predictors for each forecasting horizon.

3.1. Dataset

The data used in this study comes from the Primary Care Services of the Catalan Institute of Health (ICS), the main public health provider in Catalonia, Spain. The database is retrospective and contains all primary care visits from the period 2010-2019.

Each visit is associated with a diagnosis d_i that belongs to a set of diagnoses $D = d_1, \ldots, d_N$. In total, the database contains information on 6,301,095 patients, and the diagnoses are coded according to the medical ontology CIM-10 [13]. Only the first three letters of the diagnostic codes have been used, with the codes aggregated in the ontology until the disease group level. This reduces the amount of variables in the dataset from over 9000 to 1846. Visit information is gathered in a time period of L days, $T = t_1, \ldots, t_L$.

The initial database is transformed into a matrix $M = DxT$, where each cell M_{ij} corresponds to the number of occurrences of diagnosis d_i on day t_j. The final dimension of matrix M is 6,743,438 (N = 1846, L = 3653). Observe that the matrix is sparse since not all possible diagnoses are used in a given day. Each row of the matrix represents a time series for a given diagnosis in the period of time L.

Each time series is transformed with a rolling mean of window size = 14 to eliminate the spikes generated by the weekend effect, where there are no visits to primary care. In addition, the data is scaled to have minimum and maximum values of 0 and 1, respectively to obtain easy-to-interpret outcome metrics.

3.2. Cross-correlation analysis and collinearity effects

Given an objective diagnosis to predict d^g, the first part of the model aims to eliminate any possible collinearities present in the predictor set $D - d^g$. Initially, the model computes the Pearson correlation coefficients r_{ij} between all pairs of predictors d_i and d_j. Subsequently, correlation t-tests are performed to evaluate the probability that both predictors have a significant linear relationship [14]. The t-statistic TS_{ij} associated with each combination of predictors is calculated as follows:

$$TS_{ij} = \frac{r_{ij}\sqrt{L-2}}{\sqrt{1-r_{ij}^2}} \tag{1}$$

where L is the size of the sample, that is, the constant length diagnoses time-series. The statistic follows a t-distribution with L-2 degrees of freedom. Assuming a p-value of 0.1, pairs of variables with a TS_{ij} greater than 0.9 have a significant correlation and could therefore cause multicollinearity. To evaluate the degree of collinearity among significantly correlated variables, the algorithm uses the variance inflation factor (VIF), defined as:

$$VIF_j = \frac{1}{1-c_j^2} \tag{2}$$

Parameter c_j indicates the coefficient of determination of variable d_i regressed on the remaining predictors. The algorithm iteratively eliminates variables with the highest degree of collinearity until it is left with a subset of predictors with maximum collinearity of $max(VIF) = 20.0$ (after experimentation), considering collinearity to be non-significant [15]. Therefore, we get a set of $\Delta \subseteq D - d^g$ remaining predictors that the model can use; they do not have a significant correlation or have passed the VIF test.

3.3. Lagged MLR

Our target variable is d^g, which we aim to forecast at time t according to the minimal set of predictors $P \subseteq \Delta \subseteq D - \{d^g\}$. To that end, the following MLR model is defined:

$$model(d^g, P, t, \tau) = \alpha_0 + \sum_{i \in P} \sum_{j \in [t-\tau, t-1]} \alpha_{ij} * M_{ij} \tag{3}$$

where $t \in T$, α_0 is the constant coefficient of the model, α_j is the regression coefficient of predictor $d_i \in P$ at time j, and $\tau < L$ is the lag value applied. Note that lag value can take value 0 and therefore predict the target variable using as predictors variables in the same day.

3.4. Predictors selection

The methodology that the algorithm follows to select the predictors that form the MLR model is the forward stepwise estimation method (see Algorithm 1). Initially, the variable in P with the highest correlation coefficient with the diagnosis d^g is chosen. This predictor and the target variable d^g are fitted to a linear regression according to Equation 3. The adjusted coefficient of determination R^2 is used to evaluate the goodness of fit. It indicates the percentage of the variation explained by the added prediction with respect to the baseline model. Next, the following predictor with the highest correlation with respect to diagnoses d^g is added to the model and R^2 recomputed. To decide whether the addition of the new predictor to the model produces a significant improvement, we use the F-test statistic, defined as:

$$F1 = \frac{(SSR_2 - SSR_1)/(p_2 - p_1)}{\frac{SSE_2}{L - p_2 - p_1}} \tag{4}$$

where subindexes 1 and 2 correspond to the models with the new predictor removed or added, respectively. The variables *SSR* and *SSE* refer to the sum of squares error due to regression (variability explained by the regression) and the sum of squares error due to error (variability not explained by the regression) of the model.

The significance value marked by the F-test statistic is (p-value ≥ 0.1), indicating that the model significantly improves with the addition of the new predictor.

Sometimes, due to the forward stepwise estimation strategy, it may be found that the addition of one predictor does not show significance, but adding the next one does. For this reason, the algorithm allows for up to 3 iterations to exceed the p-value of 0.1 (assuming this 10% of possible error of significance).

Finally, among all of the models computed, the model with the best MAPE (Mean Absolute Percentage Error) is chosen. In case of a tie, the model with the fewer number of predictors is preferred (simpler model). Predictors of the best model are the ones selected.

3.5. Final model assessment

To assess the correctness of the final model multiple tests are performed to ensure it meets all required assumptions. First, the joint F-test is performed to compare the best

Algorithm 1 Predictors selection

Require: $d^g, \Delta = \{d_1, \ldots, d_{|\Delta|}\}, \tau, t$
Ensure: $P \subset \Delta$
$list\_of\_models \leftarrow null$
$counter \leftarrow 0$
$sort(\Delta, d^g, correlation)$
$P \leftarrow first(\Delta)$
$SSR_1 \leftarrow SSR(model(d^g, P, t, \tau), M)$ ▷ See Equation 4
for $\delta_i \in \Delta - first(\Delta)$ **do** ▷ Iterated according to the sort outcome
  if $counter = 3$ **then**
    break
  end if
  $model_i = model(d^g, P \cup \{\delta_i\}, t, \tau)$ ▷ See Equation 3
  $SSR_2 \leftarrow SSR(model_i, M)$
  $SSE_2 \leftarrow SSE(model_i, M)$
  $F1 \leftarrow f_{test}(SSR_1, SSR_2, SSE_2, |P|, |P \cup \{\delta_i\}|)$ ▷ See Equation 4
  append(P, $\{\delta_i\}$)
  append(list_of_models, $model_i$)
  if p-value(F1) > 0.1 **then**
    $counter = counter + 1$
  end if
  $SSR_1 \leftarrow SSR_2$
end for
$best \leftarrow minimumMAPE(list\_of\_models)$
$P \leftarrow predictors(best)$

model against the simpler linear model of the target variable (see Equation 6). This is what we call the baseline model. If the p-value was higher than a 0.05 significance level, we concluded that there is enough statistical evidence that the final model fits the observations better than the intercept-only. p represents the number of predictors used in the final model. The join F-test is computed as follows:

$$F2 = \frac{SSR}{SSE}\frac{(L-p-1)}{p} \tag{5}$$

$$baseline(d^g, t, \tau) = \beta_0 + \sum_{j \in [t-\tau, t-1]} M_{ij}, d_i = d^g \tag{6}$$

Next, Kolmogorov-Smirnov (K-S) test was conducted at a 95% confidence interval to test for normal distribution in residuals [16], and the White test for heteroscedasticity at a 95% confidence interval was applied to assess the non-constant variance of regression errors and test the assumptions of the regression model [17].

4. Results

We applied the method to compute the demand for clinical visits associated with the diagnosis of hypertension, labeled "H10" according to the CIM-10 ontology; thus $d^g =$ "$H10$". Our method is applied to make a future prediction of the target variable changing progressively the lag parameter until reaching a one-month prediction horizon,

$\tau \in [0,30]$. The system trains a model for each lagging time that can vary in predictors. This means that the model best forecasting $H10$ at horizon 0 ($\tau = 0$) may have different predictors than the one forecasting the future visits associated with $H10$ $\tau \in [1,30]$.

After applying the lagging in the matrix M (this is, shifting the target variable to the future with respect to all predictors in P accordingly to τ), the train-test splitting strategy is performed to the time series with a partition of 80% for training and 20% for testing to evaluate the forecasting. Cross-validation isn't used to preserve the structure of the time series. In this case, if the study period of the database is 10 years, the first 8 years are used for training, and the remaining 2 years are used for testing. It is worth noting that as the lag $\tau \in [1,30]$ increases in the experimentation, the training set remains the same, but the test set shortens accordingly. Since the testing period is still relatively long (730 days), the results remain robust.

The models are evaluated through the analysis of the number of predictors, the F-statistic (4) and joint F-statistic (5); and the RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error) of the training and test predictions. Both RMSE and MAPE are representations of the prediction error. RMSE presents the error as units of the variable being predicted. Therefore, since we have scaled the database within the range [0-1], RMSE addresses the error on this scale. For better interpretation, MAPE represents the error as a percentage on the scale [0-100]. Additionally, MAPE allows for comparing results between experiments, even if the test set does not have the same length.

Table 1 summarizes the best models for each lag. The algorithm achieves accurate predictions in both the on-time model ($\tau(lag) = 0$) and the future prediction models ($\tau(lag) > 0$). The number of variables that form the models ranges from 4 when predicting 9 to 12 days ahead to 25 when predicting 6 or 7 days ahead (one week). All models present a constant subset of predictor diagnoses mainly related with chronicity and elderly populations:

- M15: Primary generalized (osteo)arthrosis
- E78: Pure hypercholesterolaemia
- E03: Congenital hypothyroidism with diffuse goitre
- H90: Conductive hearing loss, bilateral
- F17: Mental and behavioural disorders due to use of tobacco

Details of the results are provided in Figure 2 to 4. Figure 2 shows the on-time model (lag value τ = 0) and Figure 3 plots the model outcome when making a prediction one month ahead (lag value τ = 30). In the case of Figure 2, the prediction is significantly better (see MAPE and RMSE results of both models in table 1), although the model also captures the general trend in the one-month prediction. The x-axis in Figures 2 and 3 represents the day, while the y-axis represents a scaled range of [0,1] for the number of visits associated with the diagnosis.

Figure 4 demonstrates the behavior of the algorithm when increasing the lagging time. The x-axis represents the prediction lag τ, while the y-axis represents the percentage of prediction error achieved by each model compared to the actual time series. The error of prediction increases from 1.2% to 2.8% when the lag is increased $\tau \in [0,30]$. The stepped effect that we see in the Figure is due to the behavior of the algorithm with the application of the forward stepwise estimation method. The algorithm does not find it worthwhile to add or remove new predictors if there is no significant increase in the lag. For small and progressive increases, the algorithm tends to maintain the best model

number variables	F1	p-value (F1)	F2	p-value (F2)	MAPE train	MAPE test	RMSE train	RMSE test	predictors (diagnoses)	LAG
9	1.67	0.20	0.87	3,51E+05	1.39	1.35	0.02	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83	0
9	0.26	0.61	-0.06	1,00E+06	1.37	1.34	0.02	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83	1
9	1.19	0.27	0.31	5,80E+05	1.48	1.43	0.02	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83	2
14	0.70	0.40	-14.32	1,00E+06	1.35	1.20	0.02	0.01	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	3
14	2.06	0.15	-22.75	1,00E+06	1.40	1.26	0.02	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	4
20	1.31	0.25	-9.88	1,00E+06	1.42	1.20	0.02	0.01	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	5
26	0.55	0.45	-15.87	1,00E+06	1.37	1.14	0.02	0.01	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	6
26	0.25	0.62	-10.77	1,00E+06	1.35	1.15	0.02	0.01	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	7
12	0.81	0.37	39.09	4,65E-04	1.67	1.53	0.02	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F32	8
5	2.27	0.13	-0.24	1,00E+06	2.01	1.67	0.03	0.02	M15,E78,E03,H90,F17	9
5	0.06	0.81	-0.29	1,00E+06	2.08	1.72	0.03	0.02	M15,E78,E03,H90,F17	10
5	0.64	0.43	1.76	1,85E+05	2.13	1.76	0.03	0.02	M15,E78,E03,H90,F17	11
5	2.70	0.10	4.73	2,98E+04	2.18	1.78	0.03	0.02	M15,E78,E03,H90,F17	12
12	1.77	0.18	-45.02	1,00E+06	1.96	1.70	0.03	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F32	13
12	2.09	0.15	-48.26	1,00E+06	1.96	1.70	0.03	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F32	14
16	0.86	0.35	-19.88	1,00E+06	1.88	1.48	0.02	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	15
16	1.98	0.16	-28.14	1,00E+06	1.97	1.50	0.03	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	16
16	2.17	0.14	-28.76	1,00E+06	2.00	1.51	0.03	0.02	M15,E78,E03,H90,F17,H40,I25,F90,Z83,M81,M23,F3...	17
8	2.54	0.11	20.38	6,62E+00	2.35	1.92	0.03	0.02	M15,E78,E03,H90,F17,H40,I25,F90	18
8	1.26	0.26	13.72	2,16E+02	2.37	1.94	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	19
8	0.58	0.45	9.12	2,55E+03	2.34	1.91	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	20
8	0.16	0.69	4.67	3,09E+04	2.33	1.90	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	21
8	0.10	0.75	3.60	5,78E+04	2.42	1.95	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	22
8	0.08	0.77	3.13	7,71E+04	2.51	1.99	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	23
8	0.08	0.78	3.03	8,20E+04	2.56	2.01	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	24
8	0.02	0.88	1.59	2,08E+05	2.60	2.02	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	25
8	0.01	0.92	-1.02	1,00E+06	2.62	2.04	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	26
8	0.05	0.82	-2.35	1,00E+06	2.58	2.03	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	27
8	0.06	0.80	-2.57	1,00E+06	2.57	2.03	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	28
8	0.00	0.97	-0.41	1,00E+06	2.65	2.08	0.03	0.03	M15,E78,E03,H90,F17,H40,I25,F90	29
8	0.04	0.84	2.10	1,47E+05	2.74	2.14	0.04	0.03	M15,E78,E03,H90,F17,H40,I25,F90	30

Table 1. Best models found for each lag value. Columns: number of predictors in best model for each lag; F1 value for best model; significance F1 best model; F2 value for best model; significance F2 for best model; MAPE train error (%); MAPE test error (%); RMSE train error; RMSE test error; predictors for best model; LAG is the time horizon in days (τ)

of the previous lag, assuming a small increase in prediction error. It is not until the increase in error is significantly high that the model changes the predictors to better adapt to the variable to be predicted (the moment when the error decreases or remains stable, as observed in the figure). This behavior can be observed in Table 1, which shows the predictors selected for the best model at each lag.

5. Discussion

Results should be further validated in clinical practice. Most of these diagnoses are highly related to an aging population, as well as hypertension, which explains the possible underlying relation detected by the model. We can assert that the five variable identified as best predictors (M15, E78, E03, H90, F17) are highly correlated with the demand for hypertension diagnoses, and monitoring these variables could help predict trends in medical visits related to hypertension six to seven days in advance with a small forecasting error.

Having models with the ability to predict future demand would be key to improving various aspects of the healthcare system. On a macro level, the system could operate much more flexibly, anticipating the need for professionals in specific areas and times,

Figure 2. On-time hypertension test prediction model ($\tau = 0$). Y-axis: Normalized amount of hypertension diagnoses in 14 days.

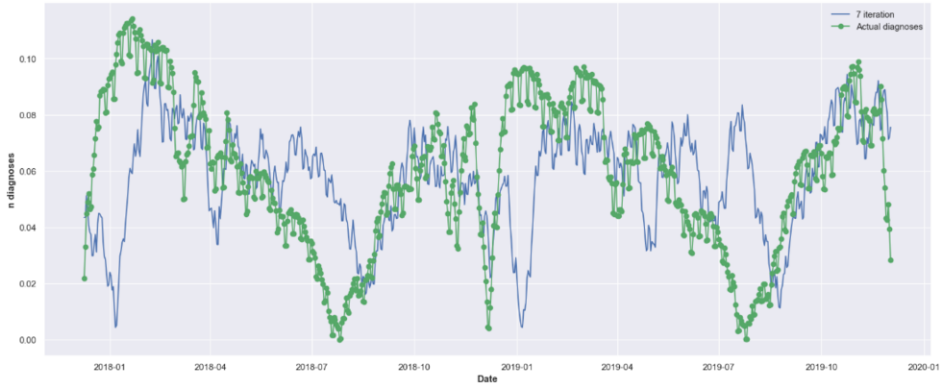

Figure 3. 30 days lagged hipertension test prediction model ($\tau = 30$). Y-axis: Normalized amount of hypertension diagnoses in 14 days.

maintaining health control over the population and also manage resources in a much more efficient way. These models would also be useful in defining disease indicators. On a micro level, healthcare professionals could have better control over their schedules by knowing what profiles they may encounter in the coming days. In a context of high demand like the one we are currently experiencing and that is expected in the future, tools like these can be key to achieving greater efficiency and flexibility in the system.

The model has strengths such as simplicity and interpretability. MLR models are widely used in the medical field and offer a high level of interpretation and acceptance by professionals. The methodology we propose can be used to select those variables most related to a target diagnosis through the feature selection method. The selection of such predictors could be a first step towards considering adding other variables related to visit diagnoses, such as medical prescriptions or pharmacological use, to improve prediction. To that end, other models like recurrent neural networks (RNN) or convolutional neural networks (CNN) could be explored.

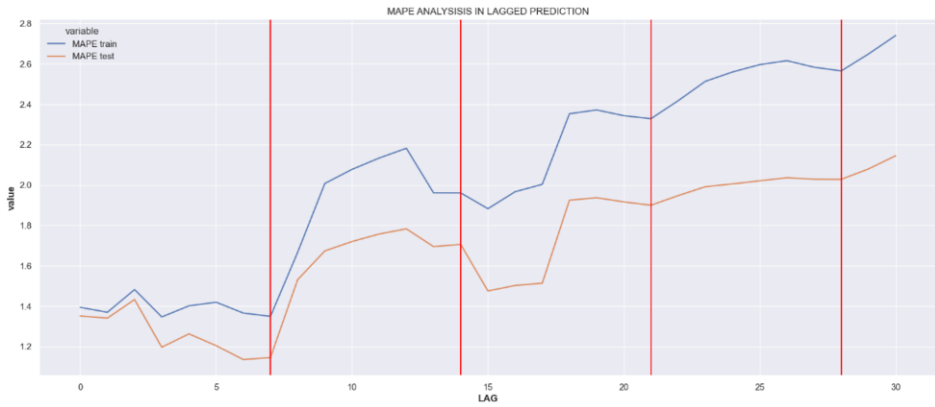

Figure 4. MAPE error lagging model.

In our work, we focus in in predicting hypertension that is one of the most common diagnostics. Despite predicting hipertension involves a quick and inexpensive test, our aim is to obtain the predictors and to understand how they behave in relation to the target variable, what their relationship is, and the time correlation that occurs between them. Moreover, other complex diagnoses as metabolic sindrome could be targered in a future.

6. Conclusions

We present in this paper a method to perform feature selection and forecast the demands of a given diagnosis (hypertension in this case) based on other diagnoses. The method is capable of indicating which predictors (other diagnostics) should be considered to predict future demand for the target, thereby increasing interpretability. In this manner, we obtain relevant information and greater explainability compared to other models, such as deep learning techniques.

The method has been tested with data of primary care services of Catalonia (ICS) in a cohort of 6,301,095 patients during years 2010-2019. The developed algorithm is capable of predicting the trend in demand for hypertension-related visits with an accuracy ranging from 98.6% to 97.86% when predicting one month in advance. Furthermore, the proposed algorithm is capable of indicating which variables are significant predictors for demand. Some of these variables have already been demostrated to have a causal relationship with hipertension in other related works [18].

Future work involves to use other variables, such as medical prescriptions, together with the predictors found; in that case, other machine learning forecasting tools, that could find some non-linearity on data, and eventually improve the forecasting results achieved by the MLR methods sould be explored.

Acknowledgements

This study was conducted with the support of the Generalitat de Catalunya, via the Industrial Doctorate Plan and via the Consolidated Research group 2021 SGR 01125.

References

[1] Lopez Segui F, Hernandez Guillamet G, Pifarré Arolas H, Marin-Gomez FX, Ruiz Comellas A, Ramirez Morros AM, et al. Characterization and Identification of Variations in Types of Primary Care Visits Before and During the COVID-19 Pandemic in Catalonia: Big Data Analysis Study. J Med Internet Res. 2021 Sep;23(9):e29622. Available from: https://www.jmir.org/2021/9/e29622.

[2] Garcia-Olive I, Lopez Seguí F, Hernandez Guillamet G, Vidal-Alaball J, Abad J, Rosell A. Impact of the COVID-19 pandemic on diagnosis of respiratory diseases in the Northern Metropolitan Area in Barcelona (Spain). Medicina Clínica. 2023;160(9):392-6. Available from: https://www.sciencedirect.com/science/article/pii/S0025775323000131.

[3] Upadhyay RK, Kumari N, Rao VSH. Modeling the spread of bird flu and predicting outbreak diversity. Nonlinear Analysis: Real World Applications. 2008;9(4):1638-48. Available from: https://www.sciencedirect.com/science/article/pii/S1468121807000879.

[4] Sil A, Kumar VN. Does weather affect the growth rate of COVID-19, a study to comprehend transmission dynamics on human health. Journal of Safety Science and Resilience. 2020;1(1):3-11. Available from: https://www.sciencedirect.com/science/article/pii/S2666449620300049.

[5] Sarkodie SA, Owusu PA. Impact of meteorological factors on COVID-19 pandemic: Evidence from top 20 countries with confirmed cases. Environmental Research. 2020;191:110101. Available from: https://www.sciencedirect.com/science/article/pii/S0013935120309981.

[6] Ahlawat A, Wiedensohler A, Mishra SK. An Overview on the Role of Relative Humidity in Airborne Transmission of SARS-CoV-2 in Indoor Environments. Aerosol and Air Quality Research. 2020;20(9):1856-61. Available from: https://doi.org/10.4209%2Faaqr.2020.06.0302.

[7] Towfiqul IAR, M H, Azad AAK, Roquia S, Zannat TF, Islam KS, et al. Effect of meteorological factors on COVID-19 cases in Bangladesh. Environment, Development and Sustainability. 2021;23(6):9139-62. Available from: https://doi.org/10.4209%2Faaqr.2020.06.0302.

[8] Morató JP, Pelegrí JL, Rey MM, Abelló AO, Vallès X, Roca J, et al. Environmental predictors of SARS-CoV-2 infection incidence in Catalonia (northwestern Mediterranean). Research Square. 2022. Available from: https://doi.org/10.21203/rs.3.rs-2206639/v1.

[9] Wang WW, Li H, Cui L, Hong X, Yan Z. Predicting Clinical Visits Using Recurrent Neural Networks and Demographic Information. 2018:353-8.

[10] Arielle S, Drake A, Emily G, L WT, Benson H, Cheryl W. Predicting unplanned medical visits among patients with diabetes: translation from machine learning to clinical implementation. BMC Med Inform Decis Mak. 2021;31;21(1):111.

[11] López B, Torrent-Fontbona F, Roman J, Inoriza JM. Forecasting of emergency department attendances in a tourist region with an operational time horizon. 2021. Available from: https://dugi-doc.udg.edu/handle/10256/19433.

[12] Spencer RJ, Amer S, George EJS. A retrospective analysis of emergency referrals and admissions to a regional neurosurgical centre 2016–2018. British Journal of Neurosurgery. 2020. Available from: https://pubmed.ncbi.nlm.nih.gov/33292027/.

[13] World Health Organization. ICD-10 : international statistical classification of diseases and related health problems: tenth revision, 2nd ed. World Health Organization https://appswhoint/iris/handle/10665/42980. 2004.

[14] Freedman D, Pisani R, Purves R. Statistics (international student edition), 4th edn. WW Norton & Company, New York. 2007.

[15] Johnston R, Jones K, Manley D. Confounding and collinearity in regression analysis - a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. Quality and Quantity. 2018 Jul;52(4):1957–1976.

[16] Jr FJM. The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association. 1951;46(253):68-78. Available from: https://www.tandfonline.com/doi/abs/10.1080/01621459.1951.10500769.

[17] White H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica. 1980;48(4):817-38. Available from: http://www.jstor.org/stable/1912934.

[18] Guillamet GH, Seguí FL, Vidal-Alaball J, López B. CauRuler: Causal irredundant association rule miner for complex patient trajectory modelling. Computers in Biology and Medicine. 2023;155:106636. Available from: https://www.sciencedirect.com/science/article/pii/S0010482523001014.