Predicting health-related conditions across different scenarios: insurance, diabetes, air quality

Marco Samorì, Simone Squaglia

Forecasting and Predictive Analytics
A.Y. 2022/2023

Introduction

This report is presented by **HealthCast**, a consulting company made up of experienced professionals and data scientists, which operates internationally. HealthCast is specialized in the health research field and works for several partners, in particular private and public, international institutions. Its purpose is to support such institutions in the deployment of their strategies and policies and in the overall improvement of their decision-making process, through the intensive use of data and predictive models. In this way, it also serves a higher goal which is to have a positive impact on the world, by exploiting the power of technology and data.

1 Health-Insurance Forecasting

1.1 Problem statement and objective

Nowadays, medical insurance is playing more and more a fundamental role with respect to people's healthcare and financial planning. In countries like the US, insurance is not provided as a public service, resulting in extremely expensive healthcare services. This forces individuals to set up health insurance coverage, in order to make medical costs more affordable and accessible. Government agencies, that are responsible for the regulation of the insurance industry, are becoming interested in predicting health insurance costs to ensure that insurers are operating in the market fairly and not charging excessive premiums on average: this is precisely what the US GOV asked us.

1.2 Data

Variable	Description	Type	Outcome
`age`	Age	Integer	
`sex`	Sex	Binary	male-female
`bmi`	Body Mass Index	Integer	
`children`	N of children	Numerical	0-5
`smoker`	Smoker/Not smoker	Binary	yes-no
`region`	Geographical region	Categorical	NE-NW-SE-SW
`charges`	Insurance charges ($)	Real	

Table 1: descriptive table of `insurance` dataset

The dataset is taken from the U.S. Census Bureau and contains 1338 observations (rows) and 7 variables (columns), about a sample from the US population which is currently enrolled in a health insurance plan, with features indicating the insured individual's characteristics and the total medical expenses charged for the calendar year. As shown above, Table 1 describes the dataset, which contains 4 numerical features (`age, bmi, children, charges`) and 3 nominal features (`sex, smoker, region`). These latter were converted into factors with numerical values, for each level.

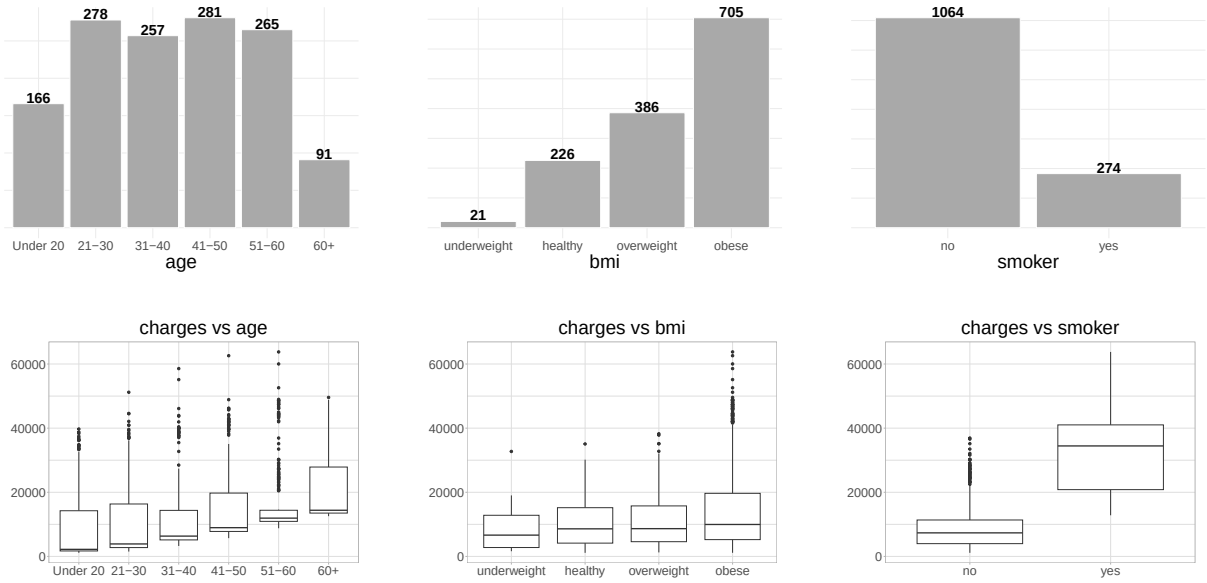


Figure 1: upper row - Barplots of : `age, bmi, smoker`;
lower row - Boxplots of : `charges` vs [`age, bmi, smoker`].

The forecast object `charges` corresponds to the health insurance plan currently paid. The distribution of `charges` is skewed on the right, indicating a strong prevalence of low insurance charges, and it presents several outliers. The exploratory data analysis highlighted the strong association of `charges` with `age`, `smoker`, `bmi`, while very low association with respect to `sex`, `children`, `region`. These relationships are shown in Figure 1 and are further confirmed by the correlation plot. As expected, being older or a smoker or not having a balanced BMI will increase the amount to be paid.

1.3 Methods

Considering the request of the US GOV, the forecasting aim is twofold: obtaining a reliable prediction of the insurance premium to be paid, alongside with an appropriate estimation of the central measure of insurance charges.

Concerning the former request, in order to obtain it, OLS had been performed to start the analysis. A comparison with log-transformation of `charges` and RLM had been explored, unfortunately without greater improvement. Further techniques were implemented to improve our prediction accuracy measure (Mean Square Error), namely: Best Subset, Ridge Regression, Random Forests.

1. Linear Regression:

$$y_i = x_i^{'}\hat{\beta} + \epsilon_i \tag{1}$$

The estimation of $\hat{\beta}$ is carried out via OLS.

In order to address the latter request of the US GOV and given the distribution of our dependent variable and the heteroskedasticity present in the data, it seemed reasonable to complement the analysis with median regression. Instead of mean estimation, a median regression is indeed a more robust measure of central tendency, less sensitive to outliers, and useful to obtain consistent estimators under broader assumptions.

The median health insurance premium to be paid is set up according to $\tau = 0.50$.

The τ quantile is computed as:

1. Linear regression:

$$\hat{q}_{\tau,i}^{LR} = x_i^{'}\hat{\beta} + \hat{\sigma}q(0.50) \tag{2}$$

2. Median regression:

$$\hat{q}_{\tau,i}^{MR} = x_i^{'}\hat{\beta}_\tau \tag{3}$$

The estimation proceeds as: $\hat{\beta}$ via OLS; $\hat{\beta}_\tau$ via MR.
The q(0.50) of Eq. 2 is computed through standardized residuals.

1.4 Empirical analysis and discussion

The sample is split into training and test sets, with 70% and 30% of the total observations respectively. The best model specification turned out to be:

$$charges_i = age_i, bmi_i, smoker_i * bmi_i \tag{4}$$

The empirical analysis starts with the performance of in-sample estimations and out-of-sample predictions of linear regression across different techniques to address the first request. Table 2 summarises the most relevant results and confirms the Ridge approach to be the most suitable one:

Model	Technique	MSE
LM	OLS*	34440438
LM	Ridge Regression	2844860
LM	RF - Boosting	27081253

Table 2: Performance Evaluation of LM across different techniques.
Note: OLS indicates the best OLS model with most relevant variables.*

Then, to address the second request, the analysis follows with in-sample estimations and out-of-sample predictions with reference to the 50th quantile. In order to evaluate prediction accuracy, the indicator variable is defined as follows:

$$I_i = \mathbb{1}_i \ (y_i < q_{\tau,i}) \tag{5}$$

If $q_{\tau,i}$ is defined as the τ_{th} conditional quantile of y_i given x_i, then I_i follows a Bernoulli distribution with τ as a parameter, in our case $\tau = 0.50$. To compute the prediction accuracy, a test of unconditional coverage is conducted via the likelihood ratio test:

$$H_0: \ E(I_i) = 0.50 \qquad H_A: \ E(I_i) \neq 0.50$$

Model	P	P-value
LR	0.4701	0.2312
MR	0.4975	0.9205

Table 3: Performance Evaluation of LR and MR.

As observed in Table 3, the predictions resulting from linear regression model and median regression model performed well: the p-value associated is not statistically significant at alpha = 0.05, which leads to accepting the null hypothesis for both. However, this is true 47% of the case for LR and 49.8% for MR. From the latter result, we tend to prefer MR for predictions.

2 Diabetes Forecasting

2.1 Problem statement and objective

Diabetes is one of the most serious and prevalent chronic diseases, which is spreading more and more in developed countries worldwide, also significantly affecting the financial burden on the economy. An early diagnosis can be extremely crucial in terms of faster lifestyle changes and more effective treatments. This is why public entities and health officials increasingly rely on predictive models as an essential tool in the prevention and early diagnosis: this is exactly what HealthCast is requested to do.

2.2 Data

Variable	Description	Type	Outcome
`Diabetes`	Diabetes	Binary	0-1
`HighBP`	High Blood Pressure	Binary	0-1
`HighChol`	High Cholesterol	Binary	0-1
`CholCheck`	Cholesterol Check	Binary	0-1
`BMI`	Body Mass Index	Integer	
`Smoker`	Smoker	Binary	0-1
`HeartDisease`	Heart Disease	Binary	0-1
`PhysActivity`	Physical Activity	Binary	0-1
`Fruits`	Fruits	Binary	0-1
`Veggies`	Veggies	Binary	0-1
`HvyAlcoholConsump`	Heavy Alcohol Consumption	Binary	0-1
`GenHlth`	General Health	Numerical	1-5
`MentHlth`	Mental Health	Numerical	0-30
`PhysHlth`	Physical Health	Numerical	0-30
`DiffWalk`	Difficulty in Walking	Binary	0-1
`Sex`	Sex	Binary	0-1
`Age`	Age	Numerical	1-13
`Education`	Education	Numerical	1-4
`Income`	Income	Numerical	1-8

Table 4: descriptive table of `diabetes` dataset

The dataset is taken from the Behavioral Risk Factor Surveillance System (BRFSS), a health-related telephone survey established in 1984, that is collected annually by the Centers for Disease Control and Prevention (CDC), in the US. For this project, the dataset used refers to the year 2015 and contains 253,680 survey responses and 22 variables. However, 19 were considered as displayed in Table 4, as a result of multiple approaches taken into account: creation of subgroups per activities, arbitrary decisions, model evaluation indicators of significance, outcomes of EDA, and literature review.

The target variable `Diabetes` has been transformed into binary while independent variables `GenHlth`, `Education`, `Income` had been dichotomized (some relevant relationships are represented in Figure 2).

2.3 Method

Considering the request of the US CDC, the aim is to predict the probability of developing diabetes connected to several socio-demographic factors, to improve early diagnosis.

Let Y_i be a r.v. taking value one if the i_{th} individual develops diabetes and zero otherwise. As $p_i = E(Y_i|X_i) \in (0,1)$, it is reasonable to use probit and logistic regression:

$$p_i = E(Y_i|X_i) = \phi(X_i^{'}\beta) \tag{6}$$

$$p_i = E(Y_i|X_i) = \frac{1}{1 + exp(X_i^{'}\beta)} \tag{7}$$

where X_i is for both a vector of socio-demographic factors for the i_{th} individual.

Having found out that data suffered from class imbalance regarding the probability of developing diabetes, additional methods have been implemented: undersampling and oversampling. The former is about reducing the number of in-samples in the majority class; the latter is about increasing the number of in-samples in the minority class. Eventually, Lasso regression had been implemented to improve our prediction accuracy through regularization of the coefficient estimates and variable selection.

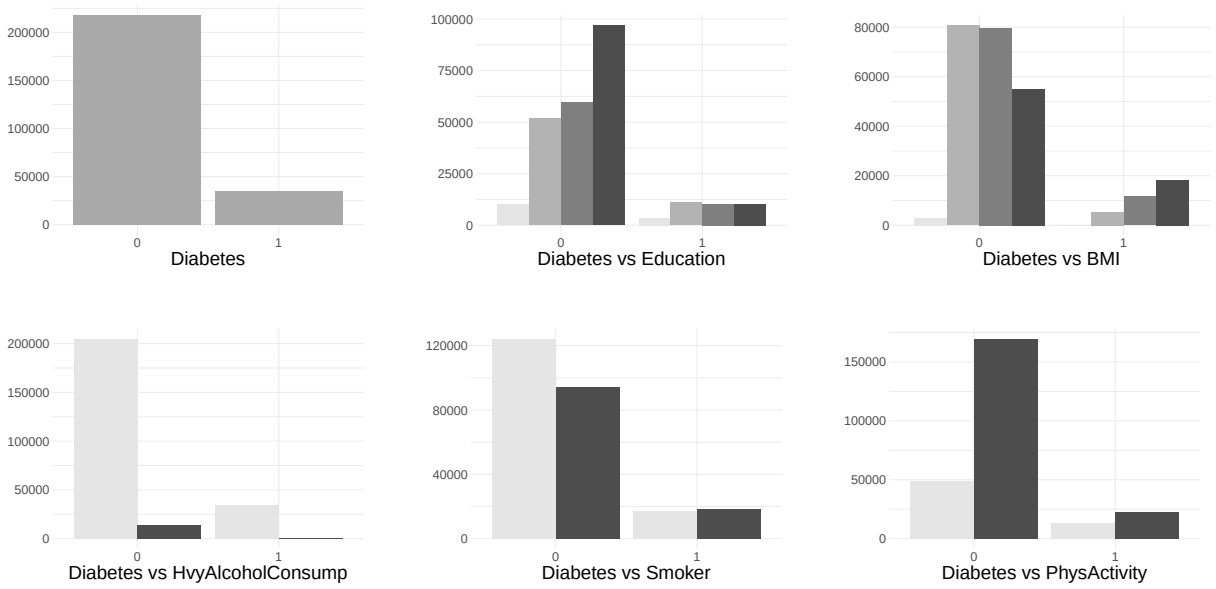


Figure 2: upper part - Barplots of : `Diabetes, Diabetes` vs [`Education, BMI`]; lower part - Barplots of : `Diabetes` vs [`HvyAlcoholConsump`, `Smoker`, `PhysActivity`].

2.4 Empirical analysis and discussion

The sample is split into training and test sets, with 70% and 30% of the total observations respectively. The empirical analysis starts with the performance of in-sample estimates and out-of-sample predictions of both probit and logistic regression. A full model has been fitted on the training set to see how it performed in its overall.

The full-model specification [Model III] is:

$$\begin{aligned} Diabetes_i = {} & HighBP_i, HighChol_i, CholCheck_i, BMI_i, Smoker_i, HeartDisease_i, \\ & PhysActivity_i, Fruits_i, Veggies_i, HvyAlcoholConsump_i, GenHlth_i, \\ & MentHlth_i, PhysHlth_i, DiffWalk_i, Sex_i, Age_i, Education_i, Income_i \end{aligned} \quad (8)$$

Adjusted-model specifications are:
[Model I] without `PhysActivity`, [Model II] without `Smoker`.

Then, it follows with out-of-sample predictions and the assessment of performance through measures based on the confusion matrix: Accuracy, Precision, Recall. In a medical setting where you're predicting a critical outcome like "Diabetes", the main appropriate metric to consider is one that focuses on minimizing false negatives (i.e., cases

where the model fails to predict diabetes, when it's actually present) and maximizing true positives (i.e., cases where the model correctly predicts diabetes). Given this trade-off, it seemed reasonable to focus primarily on the "Recall" measure:

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{TruePositive}{TruePositive + FalseNegative} \tag{9}$$

Model	Accuracy	Precision	Recall
Probit I	0.8633	0.5691	0.0937
Probit II	0.8632	0.5613	0.1012
Probit III	0.8633	0.5625	0.1014
Logit I	0.8633	0.5414	0.1485
Logit II	0.8634	0.5653	0.1004
Logit III	0.8634	0.5660	0.1002

Table 5: Performance evaluation of Probit and Logit.

As shown in Table 5, the best model turned out to be Model III for Probit and Model I for Logit. However, Recall estimates proved to be extremely low; for this reason, further techniques were carried out. At first, oversampling and undersampling were implemented in order to re-balance the `Diabetes` class but they proved to be ineffective.

Then, Lasso technique has been experimented to get more insights on the data:

Model	Accuracy	Precision	Recall
Lasso Probit	0.8485	0.3570	0.4480
Lasso Logit	0.8469	0.3666	0.4429

Table 6: Performance evaluation of Probit and Logit, using Lasso.

The results from Table 6 show a surprising increase in the Recall, at the expense of Precision measure. However, keeping in mind our objective, i.e., maximizing the Recall measure, the outcomes of the Lasso technique are quite appreciable for both Probit and Logit, with a slight preference for the Probit model.

3 Air quality Forecasting

3.1 Problem statement

Today, for most countries in the world, air pollution is becoming a major concern. The WHO has described it as "the single biggest environmental threat to human health". This is a significant shift as air quality is becoming a relevant public health risk in many countries, which are starting to rely on air quality measurements and predictions to track air pollutant concentrations and anticipate their exposure, in order to enact prompt health alerts and emergency warnings to the population. This is what the IT CNaPPS (National center for disease prevention and health promotion) requested HealthCast to do.

3.2 Data

Variable	Description	Type	Outcome
`date`	Date	Date	dd/mm/yyyy
`time`	Time	Temporal	hh.mm.ss
`CO(GT)`	CO concentration (mg/m^3)	Real	
`NMHC(GT)`	Non Metanic HydroCarbons conc (mcg/m^3)	Real	
`C6H6(GT)`	Benzene concentration (mcg/m^3)	Real	
`NOx(GT)`	NOx concentration (ppb)	Integer	
`NO2(GT)`	NO2 concentration (mcg/m^3)	Integer	
`T`	Temperature (°C)	Real	
`RH`	Relative Humidity (%)	Real	
`AH`	Absolute Humidity	Real	

Table 7: descriptive table of `airquality` dataset

The original dataset is composed of 9358 observations of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device, located on the field in a significantly polluted area, at road level, within an Italian city. Data were recorded from March 2004 to April 2005.

However, data preprocessing became imperative to address missing values, stemming from instrument malfunctions, and further dataset refinement, to be consistent with the air quality analysis. The result was a new dataset composed of 6941 observations and 10 features, as displayed in Table 7. Furthermore, temporal attributes i.e., `date` and `time`, were leveraged to derive the following columns: `DateTime`, pivotal for constructing the time series, while `Month`, `Hour`, `Day` played a crucial role in the creation of dummy variables.

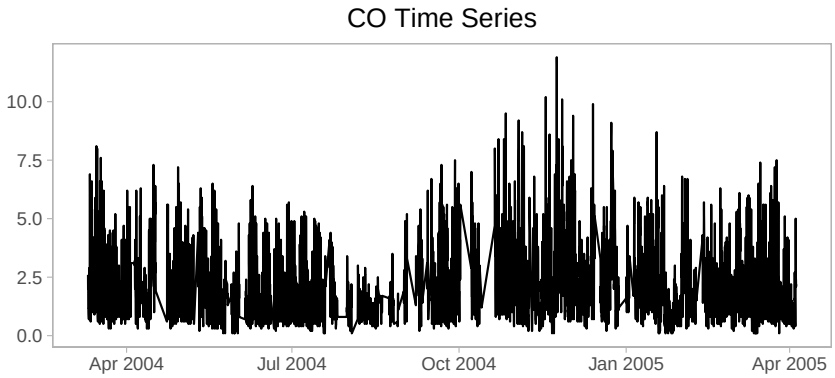


Figure 3: Time Series of `CO(GT)`.

During the Exploratory Data Analysis, the time series appeared to be stationary in mean, but not stationary in variance, as shown in Figure 3. The outcomes of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots revealed notable insights. A significant decrease in ACF values with increasing lags was observed, while the PACF indicated relevance for the first three lags. Moreover, a conspicuous seasonality pattern was evident, in particular recurring every 24 hours.

3.3 Method

The primary objective of this analysis is to forecast CO levels in a short time frame.

First of all, to assess the stationarity of the time series, the Augmented Dickey-Fuller test (ADF) was implemented:

$$H_0: \ \beta_i = 1 \qquad H_1: \ \beta_i < 1$$

The results yielded a significant finding: the time series exhibited stationarity, indicating that differencing the dependent variable was unnecessary. Therefore, in order to account for the pronounced and persistent seasonality, it sounded reasonable to construct dummy variables designed to capture and account for this effect.

At this point, the estimation proceeded by implementing different approaches; however, the most relevant ones turned out to be:

- **ARIMAX** (AutoRegressive Integrated Moving Average with eXogenous variables): approach that combines autoregressive and moving average components with external predictor variables, allowing it to capture both temporal dependencies and the influence of exogenous factors on the time series.
- **SARIMAX** (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables): extension of ARIMAX which incorporates seasonality into the model.

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{j=1}^{q} \theta_j e_{t-j} + \sum_{k=1}^{m} \gamma_k X_{k,t} + e_t \tag{10}$$

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{j=1}^{q} \theta_j e_{t-j} + \sum_{k=1}^{m} \gamma_k X_{k,t} + \sum_{s=1}^{P} \Phi_s Y_{t-s} + \sum_{s=1}^{Q} \Theta_s e_{t-s} + e_t \tag{11}$$

Subsequently, the approach followed was to delve deeper into model refinement by introducing lagged variables as potential regressors. This evolution in model complexity followed a structured approach, beginning with the inclusion of the first lags of selected variables and then extended to consider variables at lag 2.

To guide the selection of appropriate regressors, a comprehensive analysis was conducted in order to find meaningful associations that could enhance model's predictive power. Additionally, VIF was implemented for a rigorous assessment of multicollinearity and to ensure that the model remained statistically sound and interpretable.

This iterative process of model development and refinement allowed to progressively enhance the sophistication of the ARIMAX and SARIMAX models, culminating in a comprehensive understanding of the most influential variables and lags that significantly contributed to the predictive accuracy of the CO-level forecasting models.

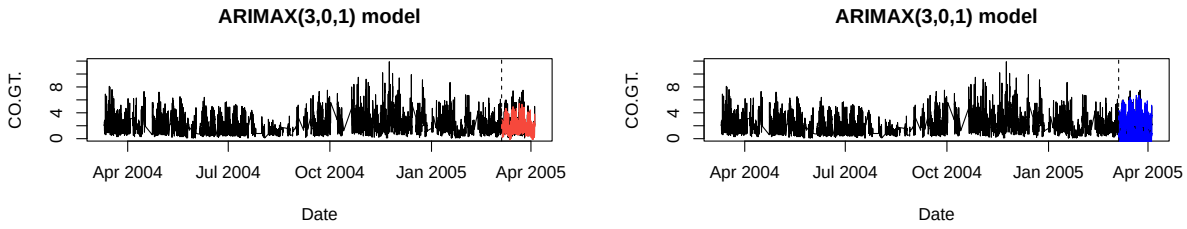
3.4 Empirical analysis and discussion

The sample is split into training and test sets, with 90% and 10% of the total observations respectively. Then, model estimation and predictions followed, with reference to a short-term scenario, where the training set was updated with the respective test values after predicting 8 steps ahead.

Model	AIC	MSE
ARIMAX(3,0,1)	12387	0.42999
SARIMAX(3,0,1)(1,0,0)[24]*	13480	0.45904

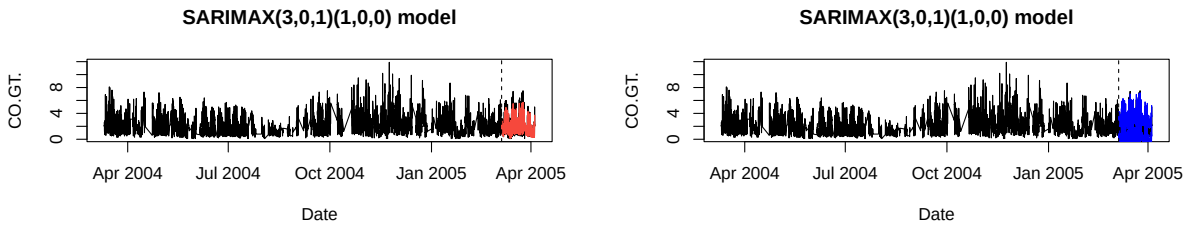
Table 8: Performance Evaluation of ARIMAX and SARIMAX.
Note: [24] indicates the period, i.e., the n. of time units in each seasonal cycle. In the case of hourly data, period = 24 indicates that the seasonality occurs every 24 hours.*

As discernible from Table 8, when assessing model performance through the AIC, the ARIMAX model yields superior results. Also, it is noteworthy that the Mean Squared Error (MSE) between the two models exhibits very little discrepancy (SARIMAX is only 6.74% higher). Given our primary objective of forecasting accuracy, it can be asserted that both models hold the potential to serve as viable alternatives within this context. A visual comparison of the predictions for the two models is shown in Figure 4 - 5.



(a) Point estimation. (b) 95% prediction interval estimation.

Figure 4: ARIMAX (3,0,1) short-term forecasting.



(a) Point estimation. (b) 95% prediction interval estimation.

Figure 5: SARIMAX (3,0,1)(1,0,0) short-term forecasting.

References

[1] James, G., Witten, D., Hastie, T., & Tibshirani, *R. (2013). An introduction to statistical learning (1st ed.)*, [PDF]. Springer.

[2] Lantz, *B. Machine Learning with R (ch. 6)*, Packt Publishing, 2013

[3] Dye, Steven. *Quantile Regression*, Towards Data Science, 2020

[4] Zanchetta, F. C., Trevisan, D. D., Apolinario, P. P., Silva, J. B., & Lima, M. H. (2016), *Clinical and sociodemographic variables associated with diabetes-related distress in patients with type 2 diabetes mellitus*. Einstein (Sao Paulo, Brazil), 14(3), 346–351. https://doi.org/10.1590/S1679-45082016AO3709

[5] Volpi, Gonzalo Ferreir. *Class Imbalance: a classification headache*, Towards Data Science, 2019

[6] Vito, Saverio (2016). *Air Quality*, UCI Machine Learning Repository. https://doi.org/10.24432/C59K5F.

[7] Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.