Data Analysis Report "Raspetout" Bank

Abderrahamane Clément
Mancer Djawed
Diot Pierre-Emmanuel
L3 Economie
2019/2020

Contents

0.1 Introduction

Our team has been assigned by **RASPETOUT Bank** to find the best model with a view to forecasting the number of customers who will get to its agency based in Tours.
We have had to set up this econometric model using a database and our knowledge in both Econometrics and R programming.
The issue we have had to resolve is actually a constrained optimization program which is to find a model with both the best forecast and the best coefficient of determination R^2 taking into account autocorrelation.
To find the model which will fit the best to the bank's data we have used a training dataset and then we have applied the model's predictions to a testing data.

0.2 First Database "mydata"

The first dataset is mostly made of time data and it reports the daily number of customers who went to the bank's agency over a 8-years-period.
The dataset has enabled us to try different models so as to find the one which perfectly fits data.

	nb <int>	t <int>	date <fctr>	jour <int>	mois <int>	an <int>	joursem <fctr>
1	462	0	02/01/2010	2	1	2010	Samedi
2	1021	0	04/01/2010	4	1	2010	Lundi
3	735	0	05/01/2010	5	1	2010	Mardi
4	604	0	06/01/2010	6	1	2010	Mercredi
5	466	0	07/01/2010	7	1	2010	Jeudi
6	434	0	08/01/2010	8	1	2010	Vendredi

Figure 1: Overview of 'mydata'

0.3 Descriptive statistics and data visualisation

After having loaded the dataset, we have manipulated the different variables so as to extract the best information from it. As we have had to deal with qualitative variables we have mainly used boxplots from the *ggplot2* package to represent data.
The different boxplots have been a good tool to highlight the days, weeks, months and other time factors which have the most influence on the number of customers at the bank agency.
This step has led us to select and create the variables we have used in several models. Here is an overview of some of the variables we have created.

	month <fctr>	year <fctr>	Lu <dbl>	Sept <dbl>	BegMonth <dbl>	EndMonth <dbl>
1	Janvier	2010	0	0	1	0
2	Janvier	2010	1	0	0	0
3	Janvier	2010	0	0	0	0
4	Janvier	2010	0	0	0	0
5	Janvier	2010	0	0	0	0
6	Janvier	2010	0	0	0	0

Figure 2: Overview of "mydatareg"

0.4 Linear regressions

In the five models we have estimated we have used both categorical variables, or factor variables, and dummy variables.

After having tested 5 linear regressions we came to a model for which $R^2 \approx 84.79\%$. Then we have corrected the model removing outlier values and correcting residuals autocorrelation.
After that, we have obtained a model with a 89.32% R^2 which means that more than 89%

of the number of customers' total variation is explained by our model. In other words, the model we have chosen replicates quite well observed outcomes.

Here is the model we have estimated to explain the quantitative variable nb representing the bank's number of customers :

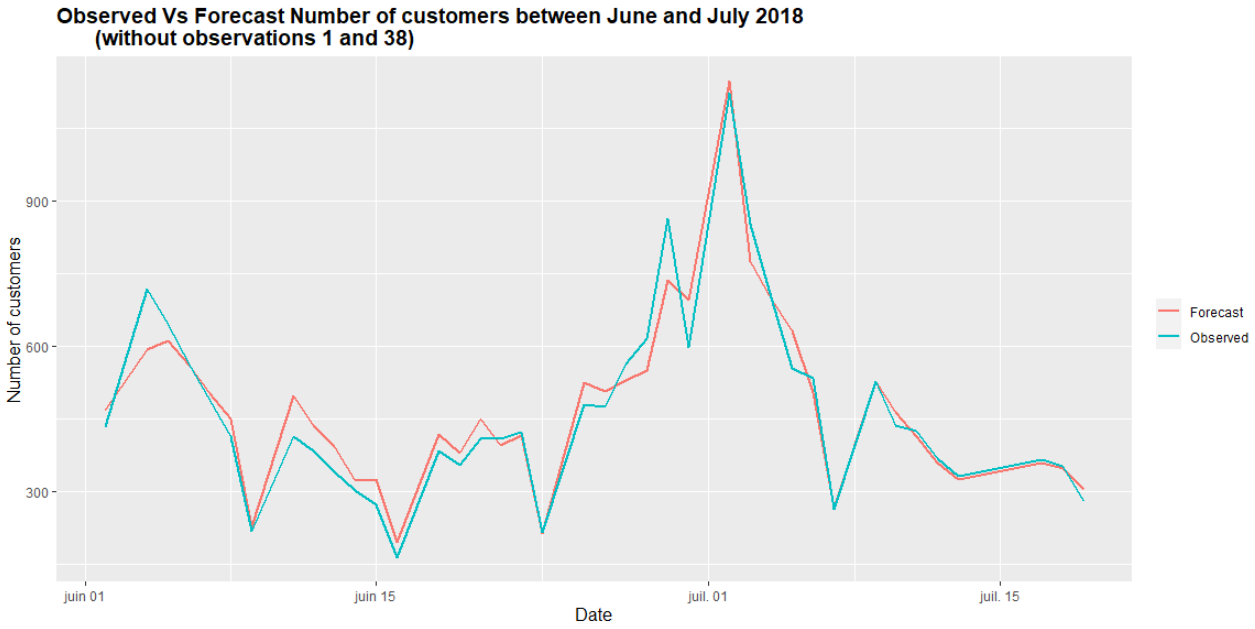
$$nb_t = \beta_0 + \beta_1 day_t + \beta_2 month_t + \beta_3 year_t + \beta_4 joursem_t + \beta_5 vacances_t + \beta_6 Lu_t \times BegMonth_t + \beta_7 BigEndMonth_t \times \beta_8 EndMonth_t + \beta_9 week_t \times MidMonth_t + \beta_{10} Sem1_t \times Sept_t + \beta_{11k}(nb-k)_t + \varepsilon_t \quad \forall t \in [\![1\ ; 2533]\!]$$

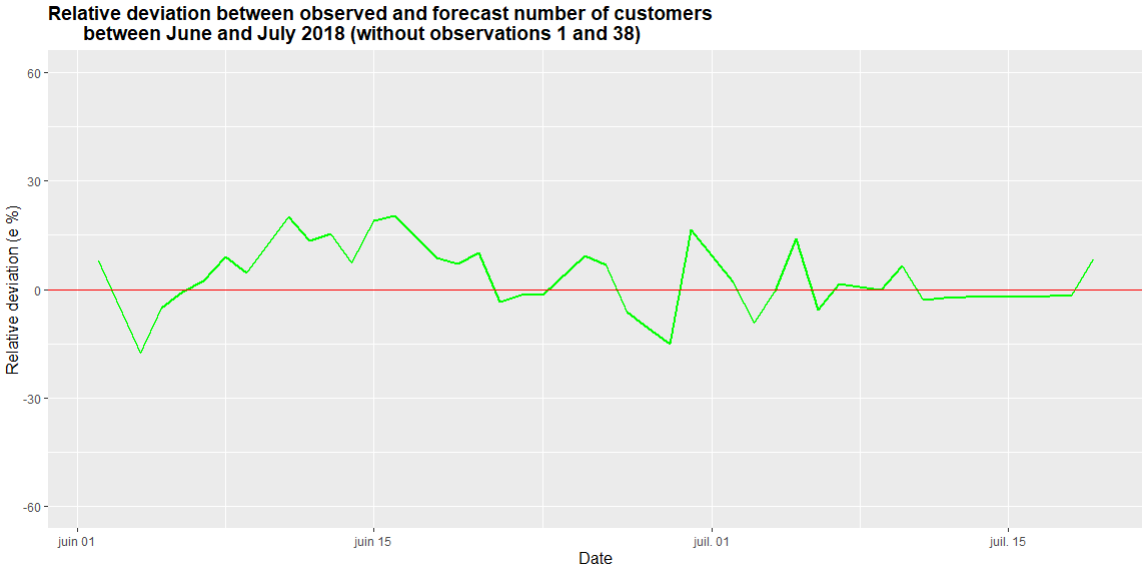
$(nb-k)_t$ represents the variable nb with a k-lag for the observation t. We have used this technique so as to solve the autocorrelation issue. The k lags stand for the residuals which are related to lagged versions of themselves.

0.5 Forecast applied to the new database 'newdata'

Eventually, we have used our model to forecast the number of customers who could get to the bank during the months of June and July 2018. We have been able to test our model on a new dataset. We have represented graphs and calculated error indicators to test whether or not our forecast would fit the observed number of customers of June and July 2018.

The following graphs show that our model fits quite well observed outcomes if we don't take into account the 1^{st} and 38^{th} observations which are outliers.





The following table depicts several econometrical error indicators.

Error Indicators about 'mymodel'

MSE	RMSE	MAE	MAPE	RSS
9720.54	98.59	50.44	0.1	3457592

MAPE is the Mean Absolute Percentage Error and it measures the accuracy of a forecasting method.
With our model we have found MAPE $\approx 10\%$ which means that on average our forecast is ten per cent smaller than the actual number of customers. We have a 10% loss of information with our forecasting method. You will find more details on the other indicators on the HTML output.