Department of Mathematics

Forecasting Categorical Time Series Using a Combination of Logistic Regression and Arima Models

Sharmina Yasmin

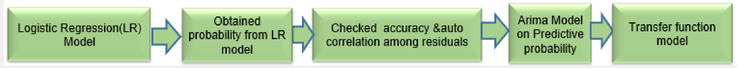
Advisor: Maher Qumsiyeh, Ph.D.

Abstract

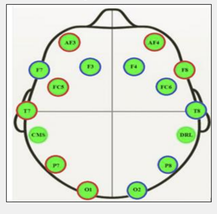
In this research, we explore a categorical time series data that changes with time and other input variables using a combination of a Logistic Regression, and ARIMA models. We use an Electroencephalogram (EEG) dataset with two states of the response variable (closed or open state of the eye). Using EEG sensor values as input, we use Logistic Regression (LR) to obtain the predictive probability to classify the eye state. Due to the time dependence, the LR model can be improved using an ARIMA model to produce better results. This will help making the residuals a white noise. This work is developed further using a Transfer Function model.

Model Development

In this work, we are studying statistical modeling with categorical time Series data. Categorical time Series data are serially correlated data for which an observation at a time step is recorded in terms of state or category.



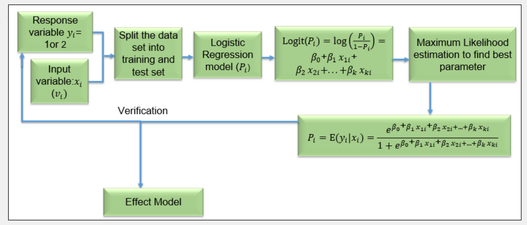
Brain wave signal-Electroencephalography (EEG), Heart pulse signal-Electrocardiography (ECG), cybersecurity, and weather data are the examples of categorical time series data. For categorical data, although the Logistic Regression (LR) model works well, however when observations are recorded sequentially over time the LR model does not work well. Due to the time dependency, the ARIMA model which works on the time series data, but it needs the numerical variable as the response variable. Thus, we developed an ARIMA model where we have used the output of the Logistic regression model which is the posterior probability obtained from LR model. Further this model is advanced with Transfer Function model with one input.



We have used an EEG data set to predict the eye state based on a given set of EEG sensor values. The figure in the left shows the sensor positions in the corpus.

Logistic Regression Model

The Logistic Regression (LR) model according to the figure, equates the Logit Transform or the log-odds of the success probability.



The maximum likelihood estimation is used to find the best parameters.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2048.6496	14	<.0001
Score	1843.7955	14	<.0001
Wald	1531.4543	14	<.0001

Small p-values from the SAS output indicate that the null hypothesis saying all the regression coefficients are zero is rejected, suggesting we have a significant model.

Accuracy of LR model

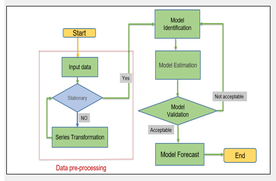
Two models are created with the first 9000 and the first 5000 observations, respectively.

	Training Model -1	Training Model-2
Number of observation	9000	5000
Test sample	3000(from end)	3000 from beginning
Test Sample		3000 from end
Model Accuracy	72.24%	73.38%
Test Accuracy	31.93%	47% (Beginning), 51% (Ending)

For model 1 and 2 the test accuracy is 31.93% and 51% respectively.

Auto correlation and ARIMA models

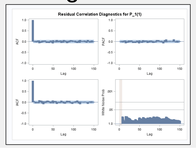
Durbin-Watson D	0.179
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	5000
1st Order Autocorrelation	0.910

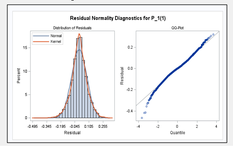


The Durbin Watson test from SAS output shows a positive autocorrelation among the residuals. According to the figure, we transform the probability of time series as stationary. Then through model identification, estimation, and validation steps a simple ARIMA model is chosen for forecasting.

Model Validation of ARIMA model

The 1st figure indicates the null hypothesis of no autocorrelation among residuals is accepted.

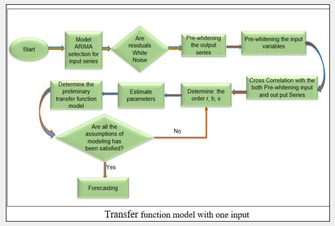




The 2nd figure indicates the residuals follow a Gaussian distribution.

Transfer Function model

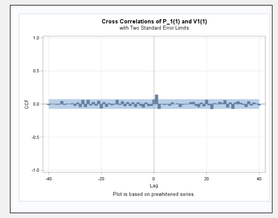
The dynamic relationship between an input and an output series is modeled by a transfer function model. The small p–value of the cross-correlation check between P1(probability) and V1(voltage) and a spike at lag one in CCF plot indicates there is a correlation between the input and the output series. The high P-values for Cross correlation check for residuals with input V1 indicates no cross correlation left between pre-whitening input and output series & the input V1 explained the output P1 as much as it can.



Transfer function model with one input

Crosscorrelation Check Between Series									
To Lag	Chi-Square	DF	Pr > ChiSq	Crosscorrelations					
5	23.03	6	0.0008	0.068	0.143	-0.074	-0.002	-0.013	-0.008
11	25.72	12	0.0118	-0.026	0.025	0.001	0.010	-0.046	-0.011
17	29.43	18	0.0434	-0.004	0.002	0.028	0.032	-0.029	-0.048
23	38.65	24	0.0297	0.021	-0.036	0.065	-0.020	-0.076	0.012
29	44.48	30	0.0431	-0.005	0.031	0.015	0.057	-0.052	0.024
35	52.18	36	0.0397	-0.074	0.024	0.041	0.018	-0.013	-0.046

Crosscorrelation Check of Residuals with Input V1									
To Lag	Chi-Square	DF	Pr > ChiSq	Crosscorrelations					
5	1.74	2	0.4191	-0.007	-0.007	0.023	-0.013	-0.022	-0.033
11	4.62	8	0.7978	0.021	0.020	0.044	-0.031	-0.011	-0.000
17	8.95	14	0.8343	-0.010	0.040	0.052	0.016	-0.035	-0.001
23	20.43	20	0.4314	-0.020	0.042	0.029	-0.094	-0.060	-0.011
29	26.32	26	0.4458	-0.006	0.002	0.063	0.004	0.023	-0.059
35	34.05	32	0.3692	-0.022	0.053	0.057	0.013	-0.047	0.039
41	38.65	38	0.4403	0.033	0.004	0.029	0.006	-0.049	0.043
47	43.29	44	0.5020	-0.045	0.034	0.020	0.008	-0.048	-0.018



Accuracy of the ARIMA Models

The two figure shows the model accuracy for ARIMA and Transfer function model with one input.

Arima Model	
Number of observation	5000
Test sample	3000(Ending)
Model Accuracy	76.7%
Test Accuracy	67.8%

Both ARIMA and Transfer function model shows 76.7% model accuracy and 67.8% test accuracy.

Transfer function model with one input	
Number of observation	5000
Test sample	3000(Ending)
Model Accuracy	76.7%
Test Accuracy	67.8%

Conclusion

Arima model improved the model and test accuracy better than the Logistic Regression model. Transfer function model with single input improved the test accuracy than the Logistic Regression model. For this work, we had 14 inputs, so we could have over 14 correlation relationship to assess in the process of figuring out which inputs are correlated with the target and how the inputs should enter the model.

Future Work

We will work on Transfer function model with multi-inputs (14 or less in our case) which might improve the test accuracy better than LR and ARIMA model.