A black background with blue and black border

Description automatically generated

**Ho Chi Minh City, May 2023**

**Lecturers:**  **Assoc. Prof. Nguyen Dinh Thuan**

**TA. Nguyen Minh Nhut**

**Students:**

**Bui Dinh Trieu - 21521576**

**Nguyen Chi Kha - 21522179**

**Nguyễn Minh Duy - 21522005**

**LAB04 REPORT**

**IS403: DATA ANALYSIS IN BUSINESS**

**VIETNAM NATIONAL UNIVERSITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

# TABLE OF CONTENTS

**Contents**

[TABLE OF CONTENTS 2](#_Toc165440706)

[TEACHER’S COMMENTS 3](#_Toc165440707)

[WORK DISTRIBUTION 4](#_Toc165440708)

[A. GAS & ELECTRIC: 5](#_Toc165440709)

[1. Using MS Excel, R, and Python Programming Language: 5](#_Toc165440710)

[1.1. Analyzing by using Excel: 5](#_Toc165440711)

[1.2. Analyzing by using Python: 12](#_Toc165440712)

[2. Conclusion: 15](#_Toc165440713)

[B. DATASET ABOUT/ OF VIETNAM: 17](#_Toc165440714)

[1. Using MS Excel, R, and Python Programming Language: 17](#_Toc165440715)

[1.1. Analyzing by using Excel: 17](#_Toc165440716)

[1.2. Analyzing by using Python: 20](#_Toc165440717)

[2. Conclusion: 23](#_Toc165440718)

[C. EVALUATE THE ACCURACY: 24](#_Toc165440719)

[Analyzing by using Python: 24](#_Toc165440720)

[Conclusion: 28](#_Toc165440721)

[D. REFERENCES: 29](#_Toc165440722)

# TEACHER’S COMMENTS

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

…………………………………………………………………………………………………

# WORK DISTRIBUTION

|  |  |  |  |
| --- | --- | --- | --- |
| **Members**  **Works** | **Nguyen Minh Duy (Leader)** | **Bui Dinh Trieu** | **Nguyen Chi Kha** |
| **Problem statement** |  |  |  |
| **Build the report template** |  |  |  |
| **Do all exercise 4.3** |  |  |  |
| **Do all exercise 4.1** |  |  |  |
| **Do all exercise 4.2** |  |  |  |
| **Summarize and edit reports** |  |  |  |
| **Completion (%)** | 100% | 100% | 100% |

## A. GAS & ELECTRIC:

**Using MS Excel and Python language to implement the ARIMA algorithm and another machine learning algorithm on the Gas & Electric dataset.**

### Using MS Excel, R, and Python Programming Language:

### Analyzing by using Excel:

**A screenshot of a graph

Description automatically generatedARIMA**

**A screenshot of a computer

Description automatically generatedA table of numbers and letters

Description automatically generated**

**Linear Regression**

A screenshot of a computer

Description automatically generated

A screenshot of a calculator

Description automatically generated

A graph with a line

Description automatically generated

A graph and chart on a white sheet

Description automatically generated

A screenshot of a computer

Description automatically generatedA graph on a white sheet

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph with blue lines and red lines

Description automatically generated

A screenshot of a computer

Description automatically generated

### Analyzing by using Python:

ARIMA

A graph of a graph

Description automatically generated

A graph of a graph

Description automatically generated

A graph of a graph

Description automatically generated

A graph with blue lines

Description automatically generated

**Linear Regression**

A screen shot of a graph

Description automatically generated

### Conclusion:

Based on the charts provided:

Gas Time Series Data:

* Both the ARIMA and AutoARIMA models predict the overall trend of the data series quite well but cannot accurately predict the peaks and troughs of the series.
* The AutoARIMA model seems to be slightly more predictive than ARIMA at capturing data series fluctuations.

Electric Time Series Data:

* For electric chains, both models have difficulty accurately predicting the peaks and troughs of the chain.
* The AutoARIMA model predicts overall trends relatively well but is unable to capture strong chain fluctuations.
* The ARIMA model seems to predict overall trends slightly better and capture fluctuations in the data series.

General comments:

* Both ARIMA and AutoARIMA are suitable models for predicting cyclic time series like this Gas and Electric data.
* However, their predictive performance depends on the nature of the data and may not be perfect at accurately predicting series peaks and troughs.
* For this particular dataset, AutoARIMA seems to give slightly better predictive results than traditional ARIMA, especially in capturing sequence fluctuations.

Based on the Linear Regression prediction charts for Gas and Electric:

* The Linear Regression model is not suitable for cyclical and seasonal time series data such as Gas and Electric:
  + Linear Regression's prediction line is a straight line, unable to capture cyclical trends and fluctuations of data.
  + This results in large prediction errors, especially at the peaks and troughs of the data series.
* Comparison with the previous ARIMA model:
  + The ARIMA model gives much better predictive results, being able to capture the cyclicality and fluctuations of Gas and Electric data.
  + ARIMA's prediction curve is more in line with actual data, especially in predicting peaks and troughs.
  + However, ARIMA is also imperfect and still has certain errors, but is still more reliable than Linear Regression for this type of data.
* Practical conclusions:
  + Linear Regression is an algorithm suitable for linear data, but not for cyclical and seasonal time series data such as Gas and Electric.
  + To predict this type of data, it is recommended to use specialized time series models such as ARIMA, SARIMA, or other machine learning models such as LSTM, Prophet, ... for more accurate prediction results.
  + Choosing the right model is critical to achieving the best predictive performance in practice.

In summary, although Linear Regression is a simple and common algorithm, it is not suitable for cyclic time series data like Gas and Electric. Other specialized models such as ARIMA will give much better predictive results on this type of data.

## DATASET ABOUT/ OF VIETNAM:

**Using MS Excel and Python language to implement the ARIMA algorithm and another machine learning algorithm on the optional dataset about/ of Vietnam**

### Using MS Excel, R, and Python Programming Language:

### Analyzing by using Excel:

**A screenshot of a computer

Description automatically generated**

A screenshot of a graph

Description automatically generated

### Analyzing by using Python:

**Dataset**

A screenshot of a computer screen

Description automatically generated

LinearRegression

A graph of a graph

Description automatically generated

Arima

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screen shot of a graph

Description automatically generated

A graph with blue dots

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated

### Conclusion:

* Higher levels of education (undergraduate, postgraduate) seem to be associated with higher levels of satisfaction with insurance service than lower levels of education (partly undergraduate).
* Insurers may consider improving service quality and communication to better meet the needs of customers with lower levels of education.
* For customers with higher education (undergraduate, postgraduate), similar policies and services can be applied because they have similar levels of satisfaction.

**These conclusions can be useful for the insurer in understanding customer needs and satisfaction based on education level, thereby adjusting business and customer service strategies accordingly.**

## EVALUATE THE ACCURACY:

**Methods to evaluate the accuracy of an algorithm or prediction model? Apply the above methods to labs 2, 3, 4.1, 4.2.**

### Analyzing by using Python:

**Lab2**

Read and check data:

A table with text and numbers

Description automatically generated A table with numbers and text

Description automatically generated

Result:

A white rectangular grid with black text

Description automatically generated

A screenshot of a spreadsheet

Description automatically generated

Using Python to evaluate the model:

Import library and data, perform data processing:

A computer code on a black background

Description automatically generated

Perform model evaluation:

A computer screen shot of a program

Description automatically generated

=> The indicators are considered acceptable

**Lab3**

Read and check data:

**A table with numbers and percentages

Description automatically generated**

Text

Description automatically generated

A number and text on a white background

Description automatically generated A number of numbers on a white background

Description automatically generated A screenshot of a computer code

Description automatically generated

=> The indicators are considered acceptable

**Lab4**

Read and check data:

A screenshot of a table

Description automatically generated

Using Python to evaluate the model:

Import library and data, perform data processing:

A black background with colorful text

Description automatically generated

Perform model evaluation:

**A screenshot of a computer

Description automatically generated**

=> The indicators are considered acceptable

### Conclusion:

**Based on the calculated model metrics, we can make the following observations:**

* Mean Absolute Error (MAE): MAE measures the average absolute deviation between the predictions and the actual values. A lower MAE indicates a more accurate model.
* Mean Squared Error (MSE): MSE measures the average squared deviation between the predictions and the actual values. Lower MSE values indicate a more accurate model.
* Root Mean Squared Error (RMSE): RMSE is the square root of MSE and is used to evaluate prediction errors on the same scale as the original data.
* R-squared (R2): R2 measures the proportion of the variance in the dependent variable that is explained by the model. R2 values closer to 1 indicate a better model fit. In this case, R2 values for both datasets are close to 0, indicating that the model does not explain the variability in the data well.

**In summary, it may be necessary to experiment with more complex prediction models or adjust factors to improve the accuracy of the predictions.**

## REFERENCES:

1. Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling (pp. 278-292). Stanford University Press.
2. Gastwirth, J. L., Gel, Y. R., & Miao, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. Statistical Science, 24(3), 343-360.
3. Kozak, M., & Piepho, H. P. (2018). What's normal anyway? Reconsideration and recommendations for the use of transformation to calculate a better estimator of the variance for quantitative variables. PloS One, 13(9), e0202976.
4. Tabachnick, B. G., & Fidell, L. S. (2019). Using Multivariate Statistics (7th ed.). Pearson Education.
5. Weiss, N. A. (2017). Introductory Statistics (10th ed.). Pearson Education.
6. Abdi, H., & Williams, L. J. (2010). Newman-Keuls test and Tukey test. Encyclopedia of Research Design, 1-11.
7. Ruxton, G. D., & Beauchamp, G. (2008). Some suggestions about appropriate use of the Kruskal-Wallis test. Animal Behaviour, 76(3), 1083-1087.
8. Shingala, M. C., & Rajyaguru, A. (2015). Comparison of post hoc tests for unequal variance. International Journal of New Technologies in Science and Engineering, 2(5), 22-33.