Cardiff Metropolitan University | Prifysgol Metropolitan Caerdydd

CIS5026_S2_20 Applied Data Science

ACTIVITY – 2 (Correlation and Regression)

Learning Outcome

1. To get acquainted with the functionalities of scikit-learn in data analysis
2. To appreciate the role of basic correlation and regression analysis in answering some interesting questions based on a dataset
3. To understand a wider role of regression analysis (example – NARMAX) for revealing the underlying dynamics and make some predictions about the output of interest
4. To Propose an interesting business analytic question based upon the available data

TASK 1

Pre-process the data – mean population of each country and mean per capita GDP (from 1995 to 2015) by making some arrangements for the missing values (HINT: You will need to use the datasets World GDP.csv and World Pop.csv). Present a correlation plot (scatter plot) between mean population of each country and mean per capita GDP (from 1995 to 2015). Very briefly interpret the generated plot. Evaluate the Pearson Correlation Coefficient.

TASK 2

Perform linear regression, where the independent variable is mean population of each country (from 1995 to 2015) and dependent variable is mean per capita GDP (from 1995 to 2015).

TASK 3

Perform a nonlinear linear regression, where the independent variable is mean population of each country (from 1995 to 2015) and dependent variable is mean per capita GDP (from 1995 to 2015). You should not aim to go beyond the degree 5 polynomial while fitting a suitable curve to the data.

TASK 4 (OPTIONAL)

Read the research article, Daily Energy Price Forecasting Using a Polynomial NARMAX Model, available on the Moodle. (a) Comment upon the possible outcome (quality of prediction), if instead of NARMAX a static regression technique is employed for the given problem. (b) Enlist three good points about the data-driven analysis and prediction presented through this article. (c) Find out at least one critical element about this article. (d) Make a power point presentation summarising your observations under the subtasks a, b, and c.

CIS5026_S2_20 Applied Data Science

Task 5

Propose an interesting business analytic question that can be answered using the given Airbnb dataset (AB_NYC_2019.csv). The proposed question should be useful for Airbnb [Note: This is also a task of your assignment sheet WRIT1]. You are recommended to email your business analytic question to the module leader (AAnupam@cardiffmet.ac.uk) within THREE working days to receive some qualitative feedback about your chosen question. Based upon the feedback you might want to either retain your question for attempting the further tasks of the assignment or you could make some necessary modifications. Since this task is a part of your assignment so the feedback at this stage will NOT be provided more than once.

.