EDHEC
MSc in Data Analytics & Artificial Intelligence

22_M2_LI_DAI_S1_CCO_6074

PYTHON PROGRAMMING
2021-2022

Final Assignment for Continuous Assessment 50% of the Grade

START DATE: 28/09/2021 12:00PM (Paris time)
DUE DATE: 12/10/2021 11:59PM (Paris time)

Professor/Lecturer: Mario HERNANDEZ-TINOCO / Christophe CROUX

INSTRUCTIONS:

- Assignment type: Individual or Group
- Type(s) of assignment file accepted: .ipynb / .py
- Number of document expected: 1
- Number of submission attempts allowed: unlimited
- Appendix(ces) provided: no
- **SafeAssign**: this assignment will be checked for plagiarism.

Final Assignment Python Programming

The Format of the Assignment:

- Can be made either individually or in groups of maximum 3 students. **IMPORTANT**: The members of the team MUST belong to the same Python Programming Group/Class. Assignments with members belonging to different Python Programming Groups/Classes WILL NOT BE ACCEPTED.
- Must be handed either in the Jupyter Notebook (.ipynb) format or as a Python (.py) format file. It is strongly suggested to use the .ipynb format.
- The assignment is an open project on Multiple Linear Regression.

The DataSet 'card_debt':

- Contains information on average credit in dollars for 400 individuals.
- Includes complete information for several quantitative (continuous and categorical predictors):

'Income' represents individual income in thousands of dollars.
'Limit' is the credit limit.
'Rating' is the individual credit rating.
'Cards' is the number of credit cards.
'Age' is in integer representing the age of the individual.
'Education' conveys the number of years of education.
'Gender' is a categorical variable taking two values: Male and Female.
'Student' is a categorical variable taking two values: Yes and No.
'Married' is a categorical variable talking two values: Yes and No.
'Ethnicity' is a categorical variable talking three values: Asian, African-American, Caucasian.
'AvgDebt' is a quantitative variable conveying the average credit card debt of individuals.

The Objectives of the Assignment:

1. Make a descriptive analysis of the variables in the dataset.
2. Create a Multiple Linear Regression Model to explain the dependent variable average credit card debt (AvgDebt)

Suggestions for the Analysis:

- 1) Divide the analysis of the summary statistics in two sections: summary statistics for quantitative (continuous variables) and categorical (qualitative) variables. Present summary statistics for the whole dataset and individual features. Summary statistics can be presented both in tables and visual graphics such as plots of histograms, correlation tables, etc. For the categorical variables: use se value counts and count plots to see the distribution of categorical values.
- 2) Use t summary statistics to make an analysis of the variables that could potentially have a high explanatory power regarding the dependent variable. E.g., use correlation tables, seaborn pairplots, scatterplots, etc. For the categorical variables, use summary

statistics such as mean, standard deviations, etc., of the dependent variable… BY GROUPS. DISCUSS which ones are the most likely to have high explanatory power in the development of the linear model. You can create regression plots of individual explanatory variables relative to the dependent variable, DISCUSS their interpretation and use it to justify their inclusion in the model.

-

3) Development of the Multivariate Linear Models: Using the Jupyter Notebook on Linear Models (last Session) as inspiration, develop a Multiple Linear Regression Model using the statsmodels package. INTERPRET the results as we did in class.

BE CREATIVE. Keep in mind that this is an open project; this is your opportunity to transform your Python skills into a concrete project. Therefore, you have the freedom to use as many techniques as you judge appropriate.

The Evaluation of the Project.

The **Evaluation** of the open project will take into account the following criteria:

- The application of the tools (libraries, methods, analyses) suggested in the minimum requirements.
- Your programming skills
- The relevance and use of the tools (libraries, methods, analyses) used for the development of the Multivariate Linear Models section.
- The quality and the level of creativity and proactivity in the development of your project.

The Submission Method and Deadline.

PLEASE INCLUDE THE NAME(S) OF THE STUDENT OR TEAM MEMBERS AT THE BEGINNING OF THE SUBMITTED FILE (JUPYTER NOTEBOOK).
The notebook must be uploaded to the assignment space created on Blackboard. The deadline to submit the project will be Tuesday October 12th at 23:59.

I hope you have fun programming in Python and that it will be very useful in the future!

Kind regards,
Mario Hernandez-Tinoco
Christophe Croux