School of Business Administration,
Economics, Law and Social Science (HSG)

Python – Heart Disease Evaluation

A description of the overall project, its individual parts,
and how to run the code

Python Code Documentation

Thomas Baer (Hubschrauber; 14-609-374)
Simon Dörpinghaus (SD_Serpent; 18-625-822)
Nicolas G. Werner (CodyCo; 14-607-816)

Dr. Mario Silic
Programming with Advanced Computer Languages
Institute of Information Management

Spring Semester 2019

A Overall Project Description

Our python programming project covers a prevalent topic in the field of medicine. While looking for insightful analyses to perform, we came across list of possible heart disease patients. The attached file includes not only key variables (such as age and chest pain), but also indicates for each patient if she/he suffered from a heart disease. Building on this data set, we decided to structure our project in three steps. First, it was our objective to analyze the heart disease data set thoroughly and display insightful findings using numerous graphs and charts. Secondly, we set the goal to develop a multiple logistic regression to predict the user's likelihood of having a heart disease by inquiring multiple inputs from the user. Thirdly, we used a second data set containing all hospitals in the United States and their corresponding quality-rankings. After having asked for the user's location (which must be in the U.S.), we intended to provide recommendations on the highest-rated hospitals in the user's state.

Overall, it was our goal to challenge ourselves by creating an original, insightful, and entertaining code. It was crucial to us that – despite the rather serious topic – we create a positive, fun(ny), and enjoyable experience for the user. We also want to emphasize that the results of the heart disease likelihood assessment are not statistically representative. While we aimed at using correct statistical methods (i.e. a multiple logistic regression), the heart disease data set is simply too small to provide accurate results.

B Individual Project Parts

I Data Evaluation and Display

Preceding the first section of our group project is a short introduction, which gives the user some information on how the program will unfold and ends with the question of whether or not he/she is ready to begin with the test.

Our project starts with a graphical illustration of the underlying heart disease data of our analysis. For this purpose, we trimmed the original dataset to only include the relevant variables and formatted the data for our purposes later on. The interaction with the users starts by asking for the type of information he/she would like to have visualized: Age related, gender related, and chest pain related information and graphs. A fourth option allows the user to skip this section and jump directly to the heart disease analysis. Once the user has selected a number from 1-3, she/he will be shown numerous charts and diagrams depicting selected aspects of the dataset. Moreover, the user can choose to see all 3 sections in any order and, additionally, decide to move on to the next section (heart disease likelihood assessment) at any time, without having to see the remaining illustrations. Once the 4th option is selected, the current section ends, and the user is introduced to part II – The heart disease analysis.

II Heart Disease Likelihood Assessment (Multiple logistic regression)

In the second phase, our project evaluates the user's risk of heart disease by performing a multiple logistic regression. Starting with a short intro, a logistic regression on the heart disease data set is performed and the user is asked to input parameters which could be used as predictors for an underlying heart disease. Finally, the logistic regression coefficients and the input data is leveraged to compute the user's likelihood of suffering a heart disease.

Like all regression analyses, the logistic regression is a predictive analysis, which is powerful when a certain probability 0<Y<1 is computed. The set of independent variables was chosen by practicability i.e. the user's access to the data inquired by the system. Hence, variables like blood pressure were excluded.

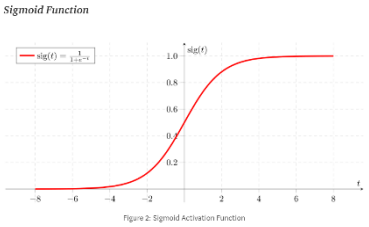

Figure 2: Sigmoid Activation Function

Disclaimer: We want to clarify that the logistic regression does not intent to be complete and scientific (with an R value lower than 0.5). The rational of this analysis was to reach a elementary level computing regressions in python and use this as a basis for further development in data analysis. Next development steps: The age seems to be quadratic, include a quadratic function for age. Additionally, multicollinearity should be tested.

III Hospital Recommendation

Thirdly, the second data set – U.S. hospital rankings – comes to use. Having evaluated the user's likelihood of a heart disease, the best possible hospitals near her/him are being recommended. To begin, we ask if the user is still OK after the "serious" heart disease likelihood assessment (and would display a funny YouTube video if not). Then, the user is asked if she/he lives in the Unites States. If not, a link to the U.S. green card application is displayed. The user can only continue once she/he moved to the United States. Living in the U.S. (either by replying yes to the original question or after having moved), the user must put in the abbreviation of her/his U.S. state. If she/he does not know the state abbreviation, a link with all state abbreviations is shown. Lastly, after having cleaned the data set, accessed those hospitals located in the user's state, and having sorted them according to their overall ranking, the user is shown a concise table of the five highest-ranked hospitals in her/his state. Subsequently, the program ends with a final "Good luck and drive safely"-message to the user.

C Running the Code

1) Download and install Anaconda using the following link: https://docs.anaconda.com/anaconda/install/windows/
2) Open the following GitHub link: https://github.com/CodyCoHSG/Hospitals
3) Download the files from GitHub
    a) HeartDisease_Project.ipynb → The code
    b) HeartDisease_Project.pdf → This file explaining the group project
4) Open the HeartDisease_Project.ipynb file with Jupyter (within Anaconda)
5) Run the code on Jupyter and follow the on-screen instructions – Enjoy!