**Dipartimento di Economia - Sede di Perugia**

Statistics for Data Science with R and Python Modulo II - Credit Scoring

**AA.2023/2024**

### Contents

Classification tools: logistic regression and discriminant analysis. These techniques will be implemented to the Credit Scoring context. Theoretical and practical notions of Credit Scoring will therefore be defined. Definition and phases; probability and independence; logistic models as classifiers; ROC and CAP curves and other validation methods (such as discriminant analysis).

### Reference texts

Alan Agresti, Maria Kateri (2021): Foundations of Statistics for Data Scientists (with R and Python). CRC Press, Chapman & Hall. ISBN: 9781003159834
Lecture notes in English (translation of the Italian book Stanghellini (2009) Introduzione ai metodi statistici per il Credit Scoring -- Springer Italia, Capp: 1-5.)

### Educational objectives

A major part of the Data Science concerns classification. Students will acquire knowledge of the major statistical methods for classification, namely logistic regression and discriminant analysis. The techniques will be applied to the Credit Scoring context, to measure the probability of default of a credit position. Real data examples and case studies through the software R and Python will give the students confidence on how to perform a data analysis in this context and learn how to build a statistical model to actually measure the risk of default. Tools presented in this module can easily applied to other contexts.

### Prerequisites

In order to successfully complete the module, students should have completed the first module Statistics for Data Science with R and Python (or any other advanced statistics course with analogous content). To be more specific: students should have successfully completed a module with (Generalized) Linear Regression covering: a) assumptions and unknown parameters; b) inferential procedures to estimate the parameters: Ordinary Least Squares, Maximum Likelihood; c) Sampling distribution of the estimators. Large sample distributions of the estimators; d) Confidence intervals. Hypothesis testing: on the parameters and on the model. F-test for the model; e) Heteroskedasticity: problems and inference in heteroskedastic models.

### Teaching methods

There will be four hours of lectures and two hours of practical exercises in the computer lab (weekly). Students are strongly advised to attend the lectures and the excercises. Furthermore, every two/three weeks, students are proposed an homework. The homework may be completed in groups of 3 or 4 students. The partecipation of the homework scheme exempt the students from providing the document 3 days prior the exams session (see Learning verification modality/Modalità di verifica dell'apprendimento below). Students are strongly advised to join the scheme.

### Other information

Incoming students within Erasmus and other Exchange programs are most welcome.

### Learning verification modality

Oral examination on both the theoretical aspects covered during the lectures and their application to real data analysis. Students are requested to complete a written report of the analysis on some given datasets, following the instructions on the file uploaded on the web page of the course in Unistudium. This document should be sent to the instructor via email three days before the exam date. Students that attend the lectures may subscribe to the programme of regular homeworks to be completed on an forthnight base. Students may do these exercises in groups. The exercises will be provided by the instructor during the lecturing time and involve solving problem on real data. This will substitute the above requested written document.

**Extended program**

Introduction to Credit Scoring as a classification method. Phases of Credit Scoring. Binary variables: odds and odds ratio. Logistic model as a generalized linear model. Interpretation of parameters. Maximum likelihood estimation of the parameters. Confidence intervals and Hypothesis testing. Classification errors. Tools for assessing the efficacy of the classifier and the accuracy of the predictors are presented, such as the ROC and CAP curves, the confusion matrix, the Hosmer-Lemeshow test. Restrospective sampling and rebalancing techniques. Discriminant analysis. Implementation of the techniques through the software R and Python for statistical computing will also be part of the course.