CREDIT SCORING – AA.2022/2023

Brief Description of the course in English. The module is taught in English. It is open to Erasmus/ incoming students. The module aims at illustrating the most used classification methods in Credit Scoring (or Rating), namely logistic regression and discriminant analysis. Tools for assessing the efficacy of the classifier, such as the ROC curve and the confusion matrix, will also be presented. Some nonparametric techniques will also be addressed..

Real data will be analized and discussed (with the softwares R and Python). For more information visit:

https://www.unipg.it/en/ects/ects-course-catalogue-2022-23?annoregolamento=2022&layout=insegnamento&idcorso=182&idinsegnamento=208472

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.

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 Scientist concerns classification. Students will acquire knowledge of the major statistic parametric techniques of classification. The techniques will be applied to the Credit Scoring Context, to measure the probability of default of a credit position. The analysis of real data and of 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 buld a statistical model to actually measure the risk of default.

Prerequisites

In order to successfully complete the module, students should have completed the module Generalized Linear Models (or any other advanced statistics course with analogous content). To be more specific: students should have successfully completed a module with Multiple 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 parameter, 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 homweork scheme exempt the students from providing the document 3 days prior the exams session (see Modalità di verifica dell'apprendimento below). Students are strongly advised to join the scheme.

Other information

Incoming students in 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 real problem on real data. This will substitute the above requested written document.

Extended program

Logistic model as a generalized linear model. Interpretation of parameters. Maximum likelihood estimation of the parameters. Confidence intervals and Hypothesis testing. Phases of Credit Scoring. 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.