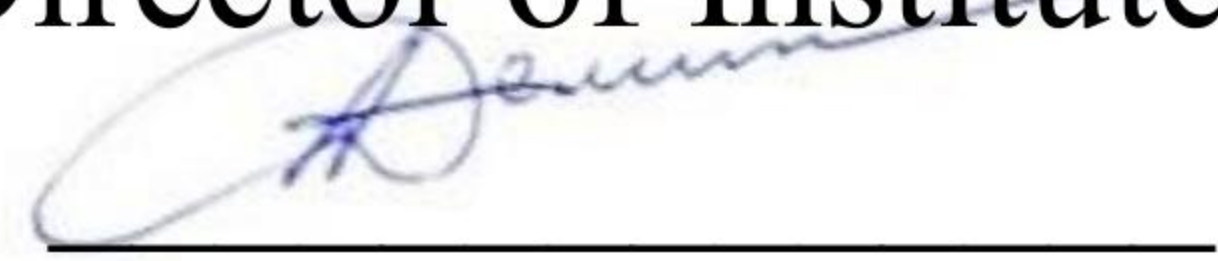


APPROVED BY

Director of Institute of Cybernetics
 / D. M. Sonkin

Special Topics in Big Data

Field of Study: Big Data Solutions

Programme name: 09.04.04 Software Engineering

Level of Study: Master Degree Programme

Year of admission: 2020

Semester, year: 4, 2

ECTS: 6

Total Hours: 216

Contact Hours: 48

- **Lectures:** 24
- **Labs:** 24
- **Practical experience:** 0

Assessment: exam, programming project

Department: Department of Software Engineering

Head of Department

 / V.S. Sherstnev

Instructor(s)

 / Gubin E.I.

Data mining on SAS platform

Course Overview

Course Objectives	The goal of this course is to introduce students to data mining on SAS platform including both the principles and techniques. Students will learn SAS Base, SAS STAT, BI Solution, and SAS Scoring for Data.
Learning Outcomes	After completing this course, the students should: • understanding the formulation of practical socio-economic problems in mathematical terms; • understand the choice of the objective function; • understand definitions of dimensions of input parameters; • should be select the forecast model; • should be evaluate of the results of a forecast model and its application in business management.
Course Outline	The course consists of 24 lectures and 24 labs covering the following main topics: The objective function; Dependent variables and their type; Test sample; A validation sample; The analysis of emissions; Regression model; Neural network; Decision tree; The ROC analysis; The Kolmogorov - Smirnov Test.
Prerequisites (if available)	Programming languages (SAS), Data Analysis Methods
Course Structure	1. The objective function. Reveals the concept of the objective function and its content. Given the basic principles of the objective function. 2. Dependent variables and their type. The definition of the dependent variable. Discusses the types of dependent variables and transformation scales. 3. Test sample. Reveals the concept of test samples, its content and volume relative to the whole volume of the source data. Given the basic principles of test samples. 4. A validation sample. Reveals the concept of the validation sample, its content and volume relative to the whole volume of the source data. Given the basic principles of test selection and validation. 5. The analysis of emissions. The concept of the audit process development. The main goals and tasks of the technological audit. 6. Checking for normality of distribution. It is supposed to test input parameters for the normality of distribution. It is necessary to use parametric statistics. 7. Test for the presence of "missingok" the source data. The presence of missing data must be replaced by a law. Either get rid of these records. 8. Regression model.

	The notion of regression models and their strengths and weaknesses. 9. Neural network. Describes concepts and the use of neural networks in forecast models. 10. Decision tree. The basic algorithms and conditions for their use. 11. The ROC analysis. The peculiarities of the given statistical criterion in assessing the quality of models 12. The Kolmogorov - Smirnov Test. For selecting the optimum cut-off score of this criterion is very useful 13. The GINI criterion. Additional statistical criterion for the optimal separation of investigated parameters.
Facilities and Equipment	3 servers with Big Data processing software (HP DL385p Gen8, 2 processors 6320 (2.8GHz-16MB) 8-Core Processor Option Kit, 6 Memory modules 8GB 2Rx4 PC3L-10600R-9 , RAID controller P420i (512MB) FBWC RAID 0,1,1+0,5,5+0, 11 HDD 500GB SC 6G 7.2K LFF SATA HotPlug Midline Drive 1y war, Flash drive 120GB 6G SATA VE 3.5in SCC EV G1 SSD) Hadoop cluster (Pig, Hive, Spark), SAS Platform
Grading Policy	In accordance with TPU rating system we use: • Current assessment which is performed on a regular basis during the semester by scoring the quality of mastering of theoretical material and the results of practical activities (labs). Max score for current assessment is 60 points, min – 30 points. • Course final assessment (exam) is performed at the end of the semester. Max score for course final assessment is 40 points, min – 25 points. • Programming project The final rating is determined by summing the points of the current assessment during the semester and exam (credit test) scores at the end of the semester. Maximum overall rating corresponds to 100 points, min pass score is 55 points.
Course Policy	Class attendance will be taken into consideration when evaluating students' participation in the course.
Teaching Aids and Resources	Compulsory Readings: 1. Credit risk modeling. Design and Application. Elizabeth Mays, Editor. Glenlake Publishing Company Ltd. 1998 2. Introduction to Scorecard for Model Builder. Fair Isaac. 2008 – 40pp. Additional Readings: 3. Scott Murray Interactive Data Visualization for the Web (http://chimera.labs.oreilly.com/books/1230000000345)
Instructor (-s)	Evgeni Gubin, +79069587250, gubine@tpu.ru