EasyChair Preprint

№ 8692

Teaching Note–Data Science Training for Finance and Risk Analysis: a Pedagogical Approach with Integrating Online Platforms

Afshin Ashofteh

August 22, 2022

Teaching Note—Data Science Training for Finance and Risk Analysis: A Pedagogical Approach with Integrating Online Platforms

Afshin Ashofteh

NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa,
Campus de Campolide, 1070-312 Lisboa, Portugal,
aashofteh@novaims.unl.pt,
WWW home page: https://novaresearch.unl.pt/en/persons/afshin-ashofteh

Abstract. The main discussion of this paper is a method of data science training, which allows responding to the complex challenges of finance. There is growing recognition of the importance of creating and deploying financial models for risk management, incorporating new data and Big Data sources, and benefiting from emerging technologies such as web technologies, remote data collection methods, user experience Platforms, and ensemble machine learning methods in finance and risk management. Automating, analyzing, and optimizing a set of complex financial systems requires a wide range of skills and competencies that are rarely taught in typical finance and econometrics courses. Adopting these technologies for financial problems necessitates new skills and knowledge about processes, quality assurance frameworks, technologies, security needs, privacy, and legal issues. This paper discusses a pedagogical approach for data science training in finance and risk analysis, with a graphical summary of necessary skills. A case study of active learning and learning by doing for financial data science course is presented, following the results of a teaching experience, online and in-person, with a combination of different technologies and platforms in an integrated manner. The outcomes of an online Q/A on the Kaggle competition platform, a book club, a video platform, and a discussion group for teaching data science for finance are presented with their advantages, disadvantages, and vulnerabilities.

Keywords: Data science, Finance, Risk, Pedagogical, Active learning

1 Introduction

Data science in finance is the analytical ability to function effectively in financial markets where data and risk are analyzed to make decisions. Data science for finance brings a range of thinking and practical skills. It includes foundations in mathematics, statistics, computer science, and finance. Moreover, the sensitivity of the outcomes to data quality needs data engineering skills [1].

In finance and risk management, the capacity to incorporate new and Big Data sources [2] and benefit from emerging technologies are investigated by many

scholars and big consultancy companies [3]. Web technologies, remote data collection techniques, user experience platforms, and blockchain brings new fields of knowledge and competencies in finance, which are necessary to automate, analyze, and optimize complex financial systems. These new scientific paradigms of information and knowledge are not included in most traditional courses in finance and econometrics [4]. It necessitates new knowledge and skills [5] and new teaching approaches to empower those thinking about a career in data science for finance to upgrade with new quality assurance frameworks, technologies, and even legislations in security, privacy, and ethical issues [6]. This implies that financial data scientists should be aware of the data protection regulations, common ethical concerns arising in financial activities, and relevant ethical guidance by national, regional, and international regulators such as IMF, central banks, and the securities and market authorities [7].

However, there are methodological issues and debates among academics concerning the many constraints to teaching the vast range of hard and soft skills and abilities considered necessary for teaching data science in finance and risk management [8]. Learning the necessary skills to use the data science solutions for financial problems effectively requires learners to be more proactive and analytically literate. Financial data scientists need to make data, methods, and outcomes more intelligible to end users and understandable in the context in which they are produced and relevant to them [9]. They should be enabled to mature the financial data analysis and apply common sense to problems to extract timely relevant information from financial data considering the uncertainty attached to them and quantify various risks. There is a need for a course in the field of financial data science to focus on these specific needs and issues relevant to financial activities and risk management.

This paper develops a framework of the essential elements for active learning of financial data science to form a meaningful learning experience. First, it presents a graphical summary in Figure 1 to show the role of data science in finance and risk management business processes. Figure 1 shows the model consists of (1) Building methodology and related theories in two design and build phases; (2) integrating methodologies with data engineering by data curators; (3) extracting the strategies based on sustainable algorithms, which are the result of combining machine learning and methodologies; (4) meta-strategy and backtesting and approval of the investment committee; (5) Graduation phase for automation and industrialization to build intelligent systems; and finally, (6) deploying the result to the platforms and checking the strategies for re-allocation if necessary. As we can see in Figure 1, soft skills such as financial thinking, data and statistical literacy, and specific knowledge of ethical codes, regulations, and dissemination of financial information are critical requirements of data science in finance.

The paper then analyses and discusses the educational requirements of this model, clarifying their contribution, interactions, and current and future importance in analysing the financial data. The learning method combines different online platforms in a harmonized way to develop soft and hard skills about data

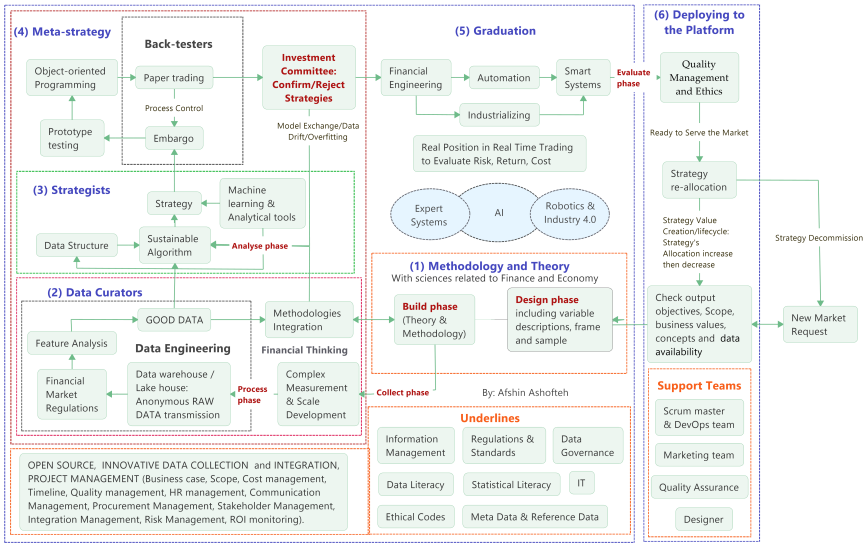

Fig. 1. Graphical summary of data science in Finance and its requirements

science for finance and its scientific paradigms. As a result, it provides information on the structure and content of a course about financial data science with an active learning process in an electronic environment (see Figure 2), which is a challenge, especially at the time of the COVID19 pandemic [10, 11].

Therefore, the remaining sections of the paper are organized as follows. In section 2, we describe the structure and content of the proposed course in financial data science. Section 3 outlines the implementation process. The results are reported and discussed in Section 4. Finally, the main conclusions are presented in section 5.

2 Structure and content of the course

The course of Financial Data Science allows learners to respond to the complex challenges of finance with new approaches to data science. It is an interdisciplinary course, and students with different backgrounds, such as Mathematics, Statistics, Computer Science, and Finance, participate in this course. Additionally, data science needs theoretical knowledge, coding skills, and soft skills such as presentation and critical thinking [12]. Following learners individually and empowering them in these areas is essential but almost impossible with the time limit of sessions and the traditional fixed design of classes. However, it would be possible if this interdisciplinary course is delivered based on problem-based learning, active learning, Learning-by-doing, and hands-on approaches by using different online platforms and technologies.

This paper presents a method with a high level of communication, presentation, and collaborative work. For this purpose, a YouTube channel, a discussion group on LinkedIn, a Kaggle project delivery platform with a Q/A, and an online programming facility were constructed. These technologies were integrated and included the videos of teacher presentations, a programming project, an online discussion group, an online troubleshooting platform, and short online articles for new topics written by the teacher with the possibility of leaving comments and sharing ideas by students, and asking students to record their presentation for one chapter of the reference book and to share them online with the possibility of putting comments by other learners. After delivering and uploading all videos of presentations, the students started to watch at least three presentations and leave brief comments about their understanding. Lecturer supervised, reviewed, evaluated, and scored all these activities, answered the questions, and corrected the wrong ideas.

The course in Financial Data Science should offer a necessary knowledge of statistical and machine learning modeling. A rigorous understanding of the modeling issues in finance is essential. It provides the tools needed to identify, measure, and manage different models' bias, variance, and error. The topics could be organized as follows:

1. Introduction to Financial data science, modeling concepts, and R/Python programming for finance.
2. Regression (Credit scoring, Simple linear regression model, Least squares criterion, Model evaluation, Multiple linear regression, Transformations, Model building, Regression pitfalls, Linear Probability Model (LPM), Logistic regression, Binary probit model) [13].
3. Time series (Time-series patterns, Trend estimation, Seasonality estimation, smoothing methods, Stationarity, Autoregressive models, Moving average models, ARMA models, Seasonal models) [14, 15].
4. Machine learning (supervised and unsupervised learning) [16].

and at the end of the semester, learners should be able to:

1. Describe the financial data science and express themselves in professional discussions.
2. Understand the importance and functioning of Regression, Time series, and Machine learning in finance.
3. Identify and distinguish the main modeling requirements and outcomes interpretation.
4. and finally, coding in R/Python for a financial problem embedded in ongoing or new work/practices and building useful reports for data-driven decisions.

3 Implementation process

This section aims to aid those thinking about teaching data science in finance, banking, and insurance.

With this in mind, the author's experience describes a course about financial data science. Furthermore, it offers some suggestions for applying different technologies to its presentation, intending to encourage relevant training in soft skills for persons involved in financial analysis and risk managers. The course is an upper-level postgraduate course, enrolling mostly juniors and seniors with different backgrounds. Because of the diverse student body, the course lecturer had a keen interest in applying different means and technologies to answer different needs. Furthermore, the learning objective listed on the syllabus shows that learners would be capable of creating and implementing advanced modeling approaches to solve financial problems. It would be possible if differentiated instructions were made to promote the diversification of materials and learning styles of different students. To obtain this end, this interdisciplinary course was preferred to be delivered based on problem-based learning, active learning, Learning-by-doing, and hands-on approaches. It tries to empower learners in both the scientific part of modeling in finance and metacognitive and socioemotional skills in a constructive learning environment. This approach needs a high level of communication, presentation, and collaborative work. Therefore, the desired educational objectives were set up; the contents were defined and organized, the proper teaching strategies were chosen for each section and topic, and the evaluation process to cover all activities was defined. Subsequently, the author revised the teaching strategies of the course to be a standalone course as much as possible.

Considering these elements was a big challenge, especially during the COVID-19 pandemic and keeping distance rules. As a result, the course was divided into three sections.

Part I of the course dealt with theoretical classes to involve students actively in the learning process. Preliminaries and slides provided an overview and information about the mechanics of the course. A forum on the moodle page of the course provided an opportunity for participants to identify themselves and say a few words about their interest in the subject of the course. Students could refer to the shared information by their classmates to choose their team members for the projects and group activities.

Part II with some support materials shared on the course Moodle page for self-study to adapt and learn more by themselves. In addition, the teacher provided supplementary handout materials in the format of articles, presentations, videos, and Q/As.

Finally, part III to make an active contribution of students in defined activities as follows:

1. Defining some small tasks and an analytical modeling project in finance and asking learners to deliver the small tasks and the final project before deadlines distributed during the semester.
2. Making a discussion group with five critical questions extracted from the course's main concepts and asking learners to answer the questions and exchange ideas. It gives students this opportunity to see the comments of their

colleagues and try to add more based on their knowledge, experience, and understanding of the theoretical classes [1].

3. Making a Kaggle competition in a teaching Kaggle profile and asking the learners to contribute to Kaggle's discussion about R/Python programming mainly related to the project of the course [2]. This activity motivates students to analyze the data, practice data handling, and work on the issues raised on modeling and machine learning concepts or complex interactions among financial data, concepts, and programming[3].
4. Building a YouTube channel with short presentation videos of the instructor to demonstrate advanced concepts and essential takeaways related to the course and project as support materials. The videos were conducted with different levels and difficulties, from basic to advance levels [4].
5. Presenting a chapter of the book: Marcos Lopez de Prado (2018), "Advances in Financial Machine Learning" by students and recording the presentations, each for a maximum of 20 minutes. Videos were shared for participants to watch at least three presentations of their colleagues, comment on divergent views on the topics according to their understanding, and raise the effectiveness of such a learning experience. It actively and extensively used the bibliography presented for this course and the chapters listed for each specific topic [5].
6. Sharing One-Pagers as a brief discussion about modeling concepts (only one page) and asking learners to read it and share their ideas in comments [6].

All these activities were supervised, reviewed, and evaluated by the lecturer. He answered the questions and corrected the wrong ideas. Each student's final grade consisted of 25% for answering the questions in the course discussion group and Kaggle, 25% for the book chapter presentation in a video format and commenting on other presentations, and 50% for the group project and problem-solving exercises. The deadlines for these activities were distributed during the semester.

In the broadest outline, the course and the underlying technology platforms are divided into six parts: Kaggle teaching facility[7], YouTube channel of the course [8], Discussion group [9], R and Python programming platform in Kaggle, [10], Data and program sharing facilities of CODEOCEAN, and Zoom online

[1] https://www.linkedin.com/feed/update/urn:li:activity:6847949183433887744?utm$_source$ = $linkedin_shareutm_medium = member_desktop_web$

[2] https://www.kaggle.com/c/credit-score-fall21/discussion/2813371581591

[3] https://www.kaggle.com/c/credit-score-fall21/overview

[4] https://www.youtube.com/channel/UCTOuxIhJxcxNOntTpamJeAA

[5] https://www.youtube.com/watch?v=I0vCTh0Sv9olist=PL$_ljXO3JR1NJXt9wD7IrwinYMP8-RCcii$

[6] https://www.linkedin.com/feed/update/urn:li:activity:6867101476813066241?utm$_source$ = $linkedin_shareutm_medium = member_desktop_web$

[7] https://www.kaggle.com/c/credit-score-fall21/overview

[8] https://www.youtube.com/channel/UCTOuxIhJxcxNOntTpamJeAA

[9] https://www.linkedin.com/groups/12420006/

[10] https://www.kaggle.com/aashofteh/code

sessions. All online materials may be copied and used for any non-commercial purpose.

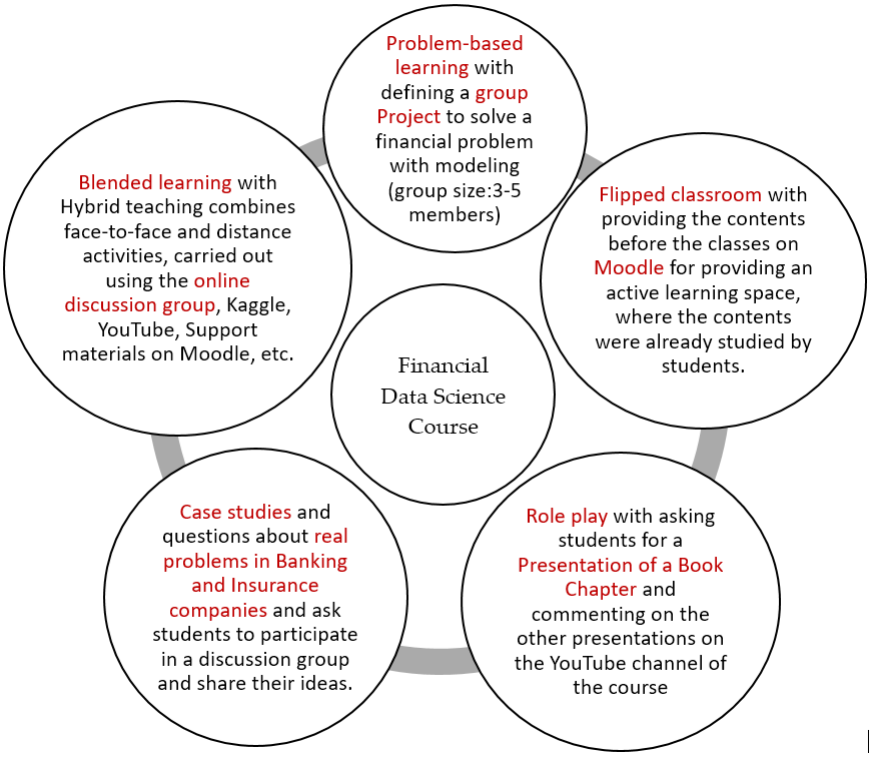

Fig. 2. Pedagogical innovation for Financial Data Science Course at the time of COVID-19 pandemic.

Figure 2 represents the graphical summary of the main activities. Although the course was designed to serve financial problems with new scientific paradigms such as data science, the discussed approach will also be valuable and exciting if adapted to other traditional courses. For instance, the author has adopted the same approach for the Banking and Insurance Operations course.

4 Results

The course evaluation shows an average of 4.5 / 5, with all individual questions higher than 4. This combination of activities was accessible online for all students. It could provide an active learning atmosphere and motivate students to participate and share their knowledge, express themselves and try to communicate and solve the group project during the semester.

Students' contributions on different platforms for different activities and learning from each other inspired them to participate actively in the learning process. Short articles provided by the teacher in the discussion group received 57 comments from students. Additionally, 250 comments from students on a discussion about the course's concepts, 15 group presentations of the reference book chapters, and 94 answers to questions about the programing project in Q/As on Kaggle that students tried to help each other, even if they were not in the same project group. This shared knowledge was produced in these activities, which is accessible online. International students had this opportunity to discuss their own country's local issues and possible solutions and share a copy of the indigenous market specifications, if it exists, along with the most recent edition of their national frameworks. Additionally, Students' contributions on different platforms for different activities and learning from each other inspired them to participate actively in the learning process. Briefly, 57 comments for One-Pager about modeling concepts, 250 comments for discussion about the course's main concepts in the discussion group, 94 discussions (Q/As) for the group project on Kaggle among students who try to help each other even if they are not in the same group, 4 shared datasets in Kaggle and 10 R or Python shared codes to help students in their projects according to their questions. Sharing codes for everyone could make it fair for all groups to benefit from the teacher's help in coding. Finally, the YouTube channel of the course included 31 videos as support materials that students could watch and leave a comment on the topics. It encouraged students to develop alternative or additional scenarios drawing on risk management problems and issues of financial markets in their own country or region.

Additionally, small group exercises were delivered to introduce some local and international issues related to risk analysis. These exercises have received attention with some relevant issues brought up by participants.

5 Main conclusions

The proposed method's innovation integrates six free online platforms for teaching a course with a reasonable workload and exact deadlines distributed in one semester. This method replaces the evaluation of students based on different activities individually and as a group project. It considers this complex evaluation instead of only the final exam, which was interesting for students. Students participated actively during the semester in different parts, worked individually on the presentations, commented on videos, answered questions of their colleagues in the discussion group, and worked as a team on their coding project. As all these activities were designed on online platforms, there was no limitation on time or place. The theoretical classes were conducted in person and as standard classes.

This approach could stimulate a vibrant and constructive learning environment at the time of COVID-19 restrictions by using a combination of different technologies and resources in the format of the text, quick Moodle quizzes,

recorded videos, online discussion groups, online Zoom trouble-shooting classes, and group projects on the Kaggle platform. It had a pretty acceptable contribution rate of students. The high number of contributions in the discussion groups, commenting on videos, and high accuracy of models constructed by students on Kaggle highlight the importance of implementing these methodologies as a pedagogical innovation in higher education to facilitate the learning process by using new technologies besides the in-person classes.

A quick study in the class about the feeling of the students about this blended method with project works, cooperative and collaborative works, and discussions shows that most students suggest these methodologies be implemented in the future.

This experience shows that the COVID-19 pandemic dramatically decreased the resistance of higher education to pedagogical innovation and provided an excellent opportunity to adapt to the structural changes that have occurred in the teaching process over the pandemic. As a result, the opportunity to use different platforms and technologies provides us with all the necessary tools to implement active learning methodologies without concerning the layout limitation of the classrooms, size of classrooms, or a high number of students.

According to this experience, there were some challenges and shortcomings. The first shortcoming of this approach was the time allotted to review and supervise all platforms and activities, including discussions, small group exercises, and responding to questions. In addition, it was time-consuming for only one presenter to monitor all these activities and evaluate them for the final grade. However, the active and friendly atmosphere of the course was the main drivers of this enjoyable experience, and available technological facilities in the classrooms motivated students to learn and implement their knowledge in practice.

Other major issues that the presenter needed to decide upon before offering this course were the number of participants with different backgrounds, and the adaptations of this multidisciplinary course to the experience and qualifications of the participants. As there is growing recognition of the importance of the data science application in finance, even some professionals in finance may participate in refresher training. As a result, the presenter should have not only strong teaching skills, leading discussions ability, and knowledge of analytical tools and programming skills [9], but also financial markets norms and regulations, with research interest related to data science and finance to be able to provide supplementary materials in elementary, intermediate, and advance levels.

The last but not least challenge is accessing financial microdata from the financial institutions of interest. In the past decade, a policy revolution has taken place among financial authorities to recognize financial microdata as confidential personal information, which should not be disseminated along with conventional publications. A good example is the ranking of financial institutions at national and international levels with the CAMELS rating system, which is confidential and may not be shared with the public, even on a lagged basis. Gathering enough real financial microdata seems necessary for practicing the phenomena

of the data science course. As a result, providing updated, anonymized, and integrated financial microdata for each chapter of the course is challenging. Even if the microdata is available, the comparability of datasets across time is a concern. Saving the same dimension and characteristics for datasets from different years is difficult, especially when we have many changes in the regulations and frameworks by internal and external financial authorities over time. The datasets must be made compatible, and it needs technical information about every influential factor on the micro dataset to ensure comparable measures are included. Thanks to some public datasets from financial institutions, one large dataset for credit risk was built for this course and shared online to the public. Access to the microdata for credit scoring example is made available at https://codeocean.com/capsule/0503126/tree/v1 at no cost.

References

1. Ashofteh, A., Bravo, J. M.: A study on the quality of novel coronavirus (COVID-19) official datasets. Stat. J. IAOS, vol. 36, no. 2, pp. 291–301 (2020). doi:10.3233/SJI-200674
2. Ashofteh, A.: Mining Big Data in statistical systems of the monetary financial institutions (MFIs). in International Conference on Advanced Research Methods and Analytics (CARMA), (2018). doi:10.4995/carma2018.2018.8570.
3. Longbing, C.: AI in Finance: Challenges, Techniques, and Opportunities. ACM Comput. Surv., vol. 55, no. 3, pp. 1–38 (2022). doi:10.1145/3502289
4. Perron, B. E., Victor, B. G., Hiltz, B. S., Ryan, J.: Teaching Note—Data Science in the MSW Curriculum: Innovating Training in Statistics and Research Methods. J. Soc. Work Educ., pp. 1–6 (2020). doi:10.1080/10437797.2020.1764891
5. Rizun, N., Nehrey, M., Volkova, N.: Data Science in Economics Education: Examples and Opportunities. pp. 550–564 (2022). doi:10.5220/0010926100003364
6. Saura, J. R., Ribeiro-Soriano, D., Palacios-Marqués, D.: Assessing behavioral data science privacy issues in government artificial intelligence deployment. Gov. Inf. Q., p. 101679 (2022). doi:10.1016/J.GIQ.2022.101679
7. Ashofteh, A., Bravo, J. M.: Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems. Stat. J. IAOS, vol. 37, no. 3, pp. 771–789 (2021). doi:10.3233/SJI-210841
8. Cahill K., et al.: Building a Computational and Data Science Workforce. jocse.org (2022). doi:10.22369/issn.2153-4136/13/1/5
9. Bonnell, J., Ogihara, M., Yesha, Y.: Challenges and Issues in Data Science Education. Computer (Long. Beach. Calif)., vol. 55, no. 2, pp. 63–66 (2022). doi:10.1109/MC.2021.3128734
10. Nacheva, R.: Emotions Mining Research Framework: Higher Education in the Pandemic Context. Contrib. to Econ., pp. 299–310 (2022). doi:10.1007/978-3-030-85254-2_18/COVER
11. Sakamaki, K., Taguri, M., Nishiuchi, H., Akimoto, Y., Koizumi, K.: Experience of distance education for project-based learning in data science. Japanese J. Stat. Data Sci., pp. 1–11 (2022). doi:10.1007/S42081-022-00154-2/TABLES/2
12. Donoho, D.: 50 Years of Data Science. vol. 26, no. 4, pp. 745–766 (2017). doi:10.1080/10618600.2017.1384734

13. Ashofteh, A.,: Big Data for Credit Risk Analysis: Efficient Machine Learning Models Using PySpark. in Proceedings of SIMSTAT 2019-10th International Workshop on Simulation and Statistics (2019).
14. Ashofteh, A., Bravo, J. M., Ayuso, M.: An ensemble learning strategy for panel time series forecasting of excess mortality during the COVID-19 pandemic. Appl. Soft Comput., vol. 128, p. 109422 (2022). doi:10.1016/j.asoc.2022.109422
15. Ashofteh, A., Bravo, J. M.: Life Table Forecasting in COVID-19 Times: An Ensemble Learning Approach. in 16th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–6 (2021). doi:10.23919/CISTI52073.2021.9476583
16. Ashofteh, A., Bravo, J. M.: A conservative approach for online credit scoring. Expert Syst. Appl., vol. 176, p. 114835 (2021). doi:10.1016/j.eswa.2021.114835