FinTech HO2020 Project

Deliverable 2.1

Grant agreement No. 825215 (Topic: ICT-35-2018 Type of action: CSA). 2021

Document Information

Deliverable information	
WP NO.	WP2
DEL. REL. NO.	D2.1
DEL. NO.	D4
TITLE	Repository of papers in big data analytics
DESCRIPTION	Repository of papers in big data analytics produced by consortium partners throughout the duration of the project.
NATURE	Document
EST. DEL. DATE	June 30 2021

Document information	
DATE	28/05/2021
WRITTEN BY	UNIPV and UBER
APPROVED BY	PAOLO GIUDICI

1

Repository of papers in big data analytics

The University of Pavia and the Humboldt University of Berlin , with the support of all the project partners, has developed the repository of papers related to Big Data Analytics. The repository contains all research papers outputted regarding big data analytics. The material has been developed by individual partners and by collaborations from within the consortium. The papers have been presented at various academic conferences and have been published in Open Access Journals or have been archived by the authors to maintain an open access copy in an Open Access Repository e.g. Arxiv, SSRN.

Specifically, the final repository contains the following information about the papers.

- Title
- Authors
- Abstract
- Partners
- Journal
- Date
- Link (doi for open access articles and also SSRN for those not open access)

Highlights

Financial technology (FinTech) solutions that employ Big Data analytics are being introduced at an extraordinary rate, particularly in credit markets, where peer-to-peer lending is emerging as a new financial service. While the application of big data analytics in peer to peer lending may reduce cost of lending, improve financial inclusion and improve user experience, it may also increase credit risks, caused by financial contagion between borrowers that derives from using a common platform.

The measurement of the additional source of credit risk due to platform contagion is of key interest to regulators and supervisors. The EU-funded FIN-TECH has developed research aimed at measuring such risk. Eleven papers have been selected for inclusion in the project's BDA research repository; three of them have been selected as use cases to be shared with regulators, supervisors, banks and fintechs. Among them, the paper "Network based scoring models to improve credit risk management in peer to peer lending platforms" has received the best feedback.

The main contribution of the paper is the proposition of a methodology based on correlation networks that can measure the contagion risk of a borrower, in terms of its network centrality, and add such risk to a classical default scoring model, based on logistic regression. The use case and the feedback from the project's stakeholders reveal that the proposed method is predictively accurate, robust, and interpretable. It can thus be suggested as a standard credit risk measurement practice for peer to peer lending applications.

Seven of the other papers selected for the project's research repository deal with the construction of credit scoring for peer to peer lending applications: employing neural network models, textual analysis, autoregressive models, spatial regression models, factor models, nearest neighbor models, ensemble models. Two papers examine the challenges and the future perspectives for peer to peer lending startups and markets, and one focuses on the broader issue of fintech risk management, from a supervisory viewpoint.

Title

Fintech Risk Management: A Research Challenge for Artificial Intelligence in Finance

Authors

Paolo Giudici

Extract

The Financial Stability Board (2017b) defines FINancial TECHnology as "technologically enabled financial innovations that could result in new business models, applications, processes, or products with an associated material effect on financial markets and institutions and on the provision of financial services."

While innovation in finance is not a new concept, the focus on technological innovations and its pace have increased significantly. Fintech solutions that make use of big data analytics, artificial intelligence and blockchain technologies are currently introduced at an unprecedented rate. These new technologies are changing the nature of the financial industry, creating many opportunities that offer a more inclusive access to financial services. The advantages notwithstanding, FinTech solutions leave the door open to many risks, that may hamper consumer protection and financial stability. Relevant examples of such risks are underestimation of creditworthiness, market risk uncompliance, fraud detection, and cyber-attacks. Indeed fintech risk management represent a central point of interest for regulatory authorities, and require research and development of novel measurements.

Partners

- University of Pavia

Journal

Frontiers in Artificial Intelligence

Data

27 November 2018

Link

https://doi.org/10.3389/frai.2018.00001

Title

Latent factor models for credit scoring in P2P systems

Authors

Daniel Felix Ahelegbey, Paolo Giudici and Branka Hadji-Misheva

Abstract

Peer-to-Peer (P2P) FinTech platforms allow cost reduction and service improvement in credit lending. However, these improvements may come at the price of a worse credit risk measurement, and this can hamper lenders and endanger the stability of a financial system. We approach the problem of credit risk for Peer-to-Peer (P2P) systems by presenting a latent factor-based classification technique to divide the population into major network communities in order to estimate a more efficient logistic model. Given a number of attributes that capture firm performances in a financial system, we adopt a latent position model which allow us to distinguish between communities of connected and not-connected firms based on the spatial position of the latent factors. We show through empirical illustration that incorporating the latent factor-based classification of firms is particularly suitable as it improves the predictive performance of P2P scoring models.

Partners

- University of Pavia

Journal

Physica A: Statistical Mechanics and its Applications

Data

10 February 2019

Link

https://doi.org/10.1016/j.physa.2019.01.130

Title

Spatial Regression Models to Improve P2P Credit Risk Management

Authors

Arianna Agosto, Paolo Giudici and Tom Leach

Abstract

Calabrese et al. (2017) have shown how binary spatial regression models can be exploited to measure contagion effects in credit risk arising from bank failures. To illustrate their methodology, the authors have employed the Bank for International Settlements' data on flows between country banking systems. Here we apply a binary spatial regression model to measure contagion effects arising from corporate failures. To derive interconnectedness measures, we use the World Input-Output Trade (WIOT) statistics between economic sectors. Our application is based on a sample of 1,185 Italian companies. We provide evidence of high levels of contagion risk, which increases the individual credit risk of each company.

Partners

- University of Pavia

Journal

Frontiers in Artificial Intelligence

Data

16 May 2019

Link

https://doi.org/10.3389/frai.2019.00006

Title

Network Based Scoring Models to Improve Credit Risk Management in Peer to Peer Lending Platforms

Authors

Paolo Giudici, Branka Hadji-Misheva and Alessandro Spelta

Abstract

Financial intermediation has changed extensively over the course of the last two decades. One of the most significant change has been the emergence of FinTech. In the context of credit services, fintech peer to peer lenders have introduced many opportunities, among which improved speed, better customer experience, and reduced costs. However, peer-to-peer lending platforms lead to higher risks, among which higher credit risk: not owned by the lenders, and systemic risks: due to the high interconnectedness among borrowers generated by the platform. This calls for new and more accurate credit risk models to protect consumers and preserve financial stability. In this paper we propose to enhance credit risk accuracy of peer-to-peer platforms by leveraging topological information embedded into similarity networks, derived from borrowers' financial information. Topological coefficients describing borrowers' importance and community structures are employed as additional explanatory variables, leading to an improved predictive performance of credit scoring models.

Partners

- University of Pavia
- Zurich University of Applied Sciences

Journal

Frontiers in Artificial Intelligence

Data

24 May 2019

Link

https://doi.org/10.3389/frai.2019.00003

Title

On the Improvement of Default Forecast Through Textual Analysis

Authors

Paola Cerchiello and Roberta Scaramozzino

Abstract

Textual analysis is a widely used methodology in several research areas. In this paper we apply textual analysis to augment the conventional set of account defaults drivers with new text based variables. Through the employment of ad hoc dictionaries and distance measures we are able to classify each account transaction into qualitative macro-categories. The aim is to classify bank account users into different client profiles and verify whether they can act as effective predictors of default through supervised classification models.

Partners

- University of Pavia

Journal

Frontiers in Artificial Intelligence

Data

07 April 2020

Link

https://doi.org/10.3389/frai.2020.00016

Title

Peer-to-peer loan acceptance and default prediction with artificial intelligence

Authors

J. D. Turiel and T. Aste

Abstract

Logistic regression (LR) and support vector machine algorithms, together with linear and nonlinear deep neural networks (DNNs), are applied to lending data in order to replicate lender acceptance of loans and predict the likelihood of default of issued loans. A two-phase model is proposed; the first phase predicts loan rejection, while the second one predicts default risk for approved loans. LR was found to be the best performer for the first phase, with test set recall macro score of 77.4%. DNNs were applied to the second phase only, where they achieved best performance, with test set recall score of 72%, for defaults. This shows that artificial intelligence can improve current credit risk models reducing the default risk of issued loans by as much as 70%. The models were also applied to loans taken for small businesses alone. The first phase of the model performs significantly better when trained on the whole dataset. Instead, the second phase performs significantly better when trained on the small business subset. This suggests a potential discrepancy between how these loans are screened and how they should be analysed in terms of default prediction.

Partners

- University College London

Journal

Royal Society Open Science

Date of Publication

10 June 2020

Link

https://doi.org/10.1098/rsos.191649

Title

Default count-based network models for credit contagion

Authors

Arianna Agosto Daniel Felix Ahelegbey

Abstract

Interconnectedness between economic institution and sectors, already recognised as a trigger of the great financial crisis in 2008–2009, is assuming growing importance in financial systems. In this article, we study contagion effects between corporate sectors using financial network models, in which the significant links are identified through conditional independence testing. While the existing financial network literature is mostly focused on Gaussian processes, our approach is based on discrete data. We indeed test dependence in the conditional mean (and volatility) of default counts in different economic sector estimated from Poisson autoregressive models, and in their shocks. Our empirical application to Italian corporate defaults in the 1996–2018 period reveals evidence of a high inter-sector vulnerability, especially at the onset of the global financial crisis in 2008 and in the following years. Many contagion effects between corporate sectors are indeed found in the shock component of the default count dynamics.

Partners

- University of Pavia

Journal

Journal of the Operational Research Society

Date of Publication

22 June 2020

Link

https://doi.org/10.1080/01605682.2020.1776169

Title

Comparing Performance of Machine Learning Algorithms for Default Risk Prediction in Peer to Peer Lending

Authors

Yanka Aleksandrova

Abstract

The purpose of this research is to evaluate several popular machine learning algorithms for credit scoring for peer to peer lending. The dataset to fit the models is extracted from the official site of Lending Club. Several models have been implemented, including single classifiers (logistic regression, decision tree, multilayer perceptron), homogeneous ensembles (XGBoost, GBM, Random Forest) and heterogeneous ensemble classifiers like Stacked Ensembles. Results show that ensemble classifiers outperform single ones with Stacked Ensemble and XGBoost being the leaders.

Partners

University of Economics - Varna

-

Journal

TEM Journal

Date of Publication

16 February 2021

Link

https://doi.org/10.1016/j.najef.2020.101318

Title

Will They Repay Their Debt? Identification of Borrowers Likely to Be Charged Off

Authors

Raluca Caplescu, Ana-Maria Panaite, Daniel Traian Pele, Vasile Alecsandru Strat

Abstract

Recent increase in P2P lending prompted for development of models to separate good and bad clients to mitigate risks both for lenders and for the platforms. The rapidly increasing body of literature provides several comparisons between various models. Among the most frequently employed ones are logistic regression, SVM, neural networks and decision tree-based ones. Among them, logistic regression has proved to be a strong candidate both because its good performance and due to its high explainability. The present paper aims to compare four pairs of models (for imbalanced and under-sampled data) meant to predict charged off clients by optimizing f1 score. We found that, if the data is balanced, Logistic Regression, both simple and with Stochastic Gradient Descent, outperforms LightGBM and K-Nearest Neighbors in optimizing f1 score. We chose this metric as it provides balance between the interests of the lenders and those of the platform. Loan term, DTI and number of accounts were found to be important positively related predictors of risk of charge off. At the other end of the spectrum, by far the strongest impact on charge off probability is that of the FICO score. The final number of features retained by the two models differs very much, because, although both models use Lasso for feature selection, Stochastic Gradient Descent Logistic Regression uses a stronger regularization. The analysis was performed using Python (numpy, pandas, sklearn and imblearn).

Partners

- Bucharest University

Journal

Management & Marketing

Date of Publication

29 August 2020

Link

https://ssrn.com/abstract=3658606

Title

Fin vs. tech: are trust and knowledge creation key ingredients in fintech start-up emergence and financing?

Authors

Theodor Florian Cojoianu, Gordon L. Clark, Andreas G. F. Hoepner, Vladimir Pažitka & Dariusz Wójcik

Abstract

We investigate how the emergence of fintech start-ups and their financing is shaped by regional knowledge creation and lack of trust in financial services incumbents across 21 OECD countries, 226 regions and over the 2007–2014 period. We find that knowledge generated in the IT sector is much more salient for fostering new fintech start-ups than knowledge generated in the financial services sector. Additionally, the importance of new knowledge created in the financial services sector (IT sector) increases (decreases) as fintech start-ups grow and seek financing. When the level of trust in financial services incumbents falls within a region, this is followed by an increase in the financing provided to fintech start-ups. Nevertheless, regions with historically low average levels of trust in financial services incumbents attract less fintech investment overall.

Partners

- University College Dublin

Journal

Small Business Economics

Date of Publication

13 June 2020

Link

https://doi.org/10.1007/s11187-020-00367-3

Title

Risk-return modelling in the p2p lending market: Trends, gaps, recommendations and future directions

Authors

Miller-Janny Ariza-Garzón; María-Del-Mar Camacho-Miñanoc; María-Jesús Segovia-Vargas; Javier Arroyo

Abstract

Peer-to-peer (P2P) lending is a market with significant growth in recent years. We review the academic literature published during the last decade on P2P lending to identify the main research trends and find potential gaps that limit stakeholders' use of research proposals. We perform both a bibliometric and systematic analysis. The bibliometric analysis will identify the most influential papers and the relationship and evolution of the main topics. In the systematic analysis, we categorized the documents according to methodological elements and business aspects. Remarkably, many proposals include artificial intelligence or machine learning algorithms. However, many of them lack a proper understanding of the application context, the definition of potential variables in a business framework, explainability, etc. Such elements should be recognized as essential elements to exploit their benefits. In this respect, we provide some recommendations and show future research directions.

Partners

- Universidad Complutense de Madrid

Journal

Electronic Commerce Research and Applications

Date of Publication

September–October 2021

Link

https://doi.org/10.1016/j.elerap.2021.101079

Annex - List of publications

Title	Authors	Journal	Keywords
Network Based Scoring Models to Improve Credit Risk Management in Peer to Peer Lending Platforms	Paolo Giudici, Branka Hadji-Misheva, Alessandro Spelta	Frontiers in Artificial Intelligence	Peer to peer lending, network based scoring models, credit risk management
Spatial Regression Models to Improve P2P Credit Risk Management	Arianna Agosto, Paolo Giudici, Tom Leach	Frontiers in Artificial Intelligence	Peer to peer lending, spatial regression models, credit risk management
Fintech Risk Management: A Research Challenge for Artificial Intelligence in Finance	Paolo Giudici	Frontiers in Artificial Intelligence	Fintech risk management, peer to peer lending, robot advisory
Will they repay their debt? Identification of borrowers likely to be charged off	Caplescu, RD., Panaite, AM., Pele, DT, Strat, VA.	Management & Marketing.	Peer to peer lending, KNN and LightGBM models, creditworthiness
Fin vs. tech: are trust and knowledge creation key ingredients in fintech start-up emergence and financing?	Theodor Florian Cojoianu, Gordon L. Clark, Andreas G. F. Hoepner, Vladimir Pažitka, Dariusz Wójcik	Small Business Economics	Fintech companies„ trust and knowledge, start-ups evolution
Peer-to-peer loan acceptance and default prediction with artificial intelligence	J. D. Turiel, T. Aste	Royal Society Open Science	Peer to peer lending, Machine learning models, loan acceptance and default prediction
On the Improvement of Default Forecast Through Textual Analysis	Paola Cerchiello, Roberta Scaramozzino	Frontiers in Artificial Intelligence	Peer to peer lending, textual analysis, credit worthiness
Latent factor models for credit scoring in P2P systems	Daniel Felix Ahelegbey, Paolo Giudici, Branka Hadji-Misheva	Physica A: Statistical Mechanics and its Applications	Peer-to-peer lending, scoring models, factor analysis models,
Default count-based network models for credit contagion	Arianna Agosto, Daniel Felix Ahelegbey	Journal of the Operational Research Society	Default prediction, financial network models, poisson autoregressive models

Comparing Performance of Machine Learning Algorithms for Default Risk Prediction in Peer to Peer Lending	Yanka Aleksandrova	TEM Journal	Peer to peer lending, Ensemble models, credit scoring
Risk-return modelling in the p2p lending market: Trends, gaps, recommendations and future directions	Miller-Janny Ariza-Garzón, María-Del-Mar Camacho-Miñano, María-Jesús Segovia-Vargas, Javier Arroyo	Electronic Commerce Research and Applications	Peer to peer lending, risk-return modelling, future evolution