Statistics and Probability Letters 136 (2018) 160–164

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Financial data science

Paolo Giudici

University of Pavia, Italy

ARTICLE INFO

Article history:
Available online 27 February 2018

JEL classification:
G01
C58
C63

Keywords:
Data science
Financial technologies
Graphical models
Network models

ABSTRACT

Data science can be defined as the interaction between computer programming, statistical learning, and one of the many possible domains where it can be applied. In the paper we provide a description of Financial data science, which involves the application of data science to technologically enabled financial innovations (FinTech), often driven by data science itself. We show that one of the most important data science models, correlation networks, can play a significant role in the advancements of Fintech developments.

1. Introduction

Data science updates the concept of data mining in the light of the availability of big data, that differ from data by their automatic generation through social networks, sensors and other data generating tools.

In this sense, data science can be defined extending Giudici (2003), as 'an integrated process that consists of a series of activities that go from the definition of the objectives of the analysis, to the selection and processing of the data to be analysed, to the statistical modelling and summary such data and, finally, to the interpretation and evaluation of the obtained statistical measures'.

This definition clearly distinguishes data science from applied statistics, or from machine learning, which concern only one part of the data science process. Data mining, on the other hand, closely resembles data science, with the latter characterised by the presence of 'big data', automatically generated.

From the previous discussion it clearly appears that the process of data science depends on what the objectives of the analysis are. Different knowledge domains may have very different objectives, expressed in very different languages. For this reason, in my opinion, it is better to talk about 'data scienceS' rather than 'data science'.

In this paper we focus on one of such data scienceS, which is known as 'Financial data science' or, in a more modern terminology 'Fintech data science'.

In the recent years, the emergence of financial technology ventures ('Fintechs') in finance has introduced many opportunities for both companies and users, which have redefined the roles of traditional financial intermediaries. Since 2005, the growth of Fintech investments has been exponential, with their total funding jumping from around $5.5B in 2005 to more than $100.2B in 2017.

A common trait of many Fintechs is that they are based on peer-to-peer (P2P) financial transactions, which give raise to service 'disintermediation' and to the need of a new protection of consumers and investors (see for example Guo et al., 2016).

Many factors can explain the increasing role of P2P platforms in finance. The first reason is that these online marketplaces can avoid intermediation costs typically associated with traditionalfinancial services. For instance, P2P platforms are not

E-mail address: giudici@unipv.it.

https://doi.org/10.1016/j.spl.2018.02.024
0167-7152/

required to respect bank capital requirements nor to pay fees associated with state deposit insurance practices, and this allows them to operate with lower costs.

A second reason is that advancements in data science have been a key force driving the exponential growth of P2P platforms. Big data analytics has changed how data is collected, processed, and evaluated, which in turn has led to significant reductions in search costs for credit information.

A third reason is that, in some cases, regulation factors may favour P2P platforms. For instance, the new European Revised Payment Service Directive (PSD2), that is going to be effective in 2018, may improve the usage of advanced data analytics.

However, the growth of marketplace platforms does not bring only advantages. It can also pose significant risks to the financial system. For example, an increased volume of lending, which brings a higher commission revenue, could be associated with the underestimation of the credit risk of the counterparty. This does not mean that P2P should be prohibited, or restricted, but that they should be motivated to make a better use of the big data they 'natively' have, as shown by Giudici and Hadji-Misheva (2017) in a recent paper.

Classical banks have, over the years, segmented their reference markets into specific territorial areas, or in specific business activities, thus increasing their expertise and, consequently, the accuracy of their ratings. Differently, P2P platforms are based on a "universal" banking model, that is fully inclusive, without space and business type limitations implying that developing an accurate rating model is a more difficult task. However, P2P platforms have the advantage of an improved data collection on the network to which a borrower belongs. A network which can be derived from the business and/or social relationships that exist among clients in the P2P network.

In line with the previous insights, Giudici and Hadji-Misheva (2017) suggest to improve the predictive performance of risk measurement models employed by P2P platforms by means of correlation network models. Here we report the main findings from their study.

Correlation networks have been recently introduced in finance, as a tool for the purpose of understanding financial flows in global markets as they apply to the management of systemic risk. See, for instance, the papers by Chinazzi et al. (2012) and Giudici and Spelta (2016). The next section recalls the main characteristics of correlation network models in finance.

2. Network models in finance

Statistical theory offers a great variety of models for predictive purposes in finance. When the quantity to be predicted is a default event (as in credit risk measurement) the most used approaches are based on logistic regression models.

Logistic regression aims to classify statistical units in a dependent variable that consists of two groups, characterised by a different loan status (1=default; 0=no default), according to the following model:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \sum_j \beta_j x_{ij},$$

where p_i is the probability of default, for borrower i, $x_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{ij})$ is a vector of borrower-specific explanatory variables, and the intercept parameter α, as well as the regression coefficients β_j, for $j = 1, \ldots, J$, are to be estimated from the available data.

The two main issues that arise in employing logistic regression for financial predictions are: (1) the event to be predicted may be rare, so there is not enough information for efficient statistical estimates; (2) the event to be predicted is multivariate, and there are strong dependences among the individual default variables.

In the case of low default frequencies, extreme value Generalised models could be used. For example, logistic regression could be extended with the Generalized Extreme regression scoring model proposed by Calabrese and Giudici (2015).

Here we focus our attention on the prediction of multiple default events, correlated among each other. This is indeed one area where data science and, in particular, network based models, can be very helpful. Lauritzen (1996) contains a very clear and summarised description of multivariate graphical models and probabilistic expert systems, employed in the '90s to model dependencies between random variables, by means of a unifying and powerful concept of a mapping between probabilistic conditional independences, missing edges in a graphical representation, and suitable model parametrisations. Two of the main limitations of the graphical modelling approach are its computational complexity, an exponential function of the number of nodes, and the presence of many competing models, which may induce a strong variability in the estimates, due to model uncertainty. While the first issue is mitigated by the increasing computational power available, the second has been tackled by Bayesian graphical modelling, as in the papers Madigan and Raftery (1994), Giudici and Green (1999), Castelo and Giudici (2001), Giudici and Stanghellini (1999), Brooks et al. (2003). These models have been applied in a variety of financial contexts, including credit scoring, churn modelling and cross selling basket analysis (for a review see Giudici, 2003).

The emergence of social networks, and of big data generated by them, has suggested the need to build graphical models whose nodes are associated with units rather than by their descriptor variables, as in the P2P running example presented here, with a noticeable increase in the number of nodes and, therefore, in model computational complexity. Accordingly, graphical models have been replaced by social network analysis, whose emphasis is on concepts such as node centrality, and interaction between nodes.

Recently, social network analysis has become increasingly recognised as a powerful methodology for investigating and modelling financial interactions between economic agents, particularly in the wake of the recent financial crisis, which has led to an increasing research literature on systemic risk, with different definitions and measurement models. According to

the European Central Bank: "Systemic risk is the risk of experiencing a strong systemic event, which adversely affects a number of systemically important intermediaries or markets". This definition introduces two key elements for the study of systemic risk: first, a trigger event, transmitted to the system as a whole, and not only to individual institutions; second, financial instability, which is spread recursively through contagion from each financial institution to another. While systemic risk definitions share this broad view, systemic risk measurement models are still quite divergent.

Specific measures of systemic risk have been proposed for the banking sector; in particular, by Acharya et al. (2012), Adrian and Brunnermeier (2016), Brownlees and Engle (2012), Acharya et al. (2012) who, on the basis of market share prices, calculate the quantiles of the estimated loss probability distribution of a bank, conditional on the occurrence of an extreme event in the financial market. The above approach is useful to establish policy thresholds aimed, in particular, at identifying the most systemic institutions. However, it is a bivariate approach, which allows to calculate the risk of an institution conditional on another or on a reference market but, on the other hand, it does not address the issue of how risks are transmitted between different institutions, in a multivariate framework. Trying to address the multivariate nature of systemic risk, researchers have recently proposed correlation network models, that combine the rich structure of financial networks (see, e.g., Lorenz et al., 2009; Battiston et al., 2012) with a parsimonious approach based on the dependence structure among market prices. The first contributions in this framework are Billio et al. (2011) and Diebold and Yilmaz (2014), who propose measures of connectedness based on Granger-causality tests and variance decompositions. Giudici and Spelta (2016) and Ahelegbey et al., (2015) took a further step, linking econometric financial networks to graphical Gaussian models, proposing the so-called 'correlation network models', subsequently extended to mixtures by Abedifar et al. (2017), Cerchiello et al. (2017), Giudici et al. (in press) and Giudici and Parisi (2017).

In the context of our study, the statistical units correspond to borrowers that are small and medium enterprises, for which correlation networks can be built on the basis of an observed balance sheet variable. Specifically, if we consider each borrower to be a node in the network and we associate each node with the series of values observed in time for the balance sheet variable, each pair of nodes can be thought to be connected by an edge, whose weight is equal to the correlation coefficient between the two corresponding series of values. More formally, a network between N borrowers can be represented by the associated matrix of weights, W, named adjacency matrix, with elements w_{xy} defined by:

$$w_{xy} = \frac{T(\sum_t x_t y_t) - (\sum_t x_t)(\sum_t y_t)}{\sqrt{[T\sum_t x_t^2 - (\sum_t x_t)^2][T\sum_t y_t - (\sum_t x_t)^2]}},$$

where $x = (x_1, \ldots, x_T)$ and $y = (y_1, \ldots, y_T)$ are the two series of values of the balance sheet variable observed, respectively, for borrowers x and y, at times $t = 1, \ldots, T$.

Network centrality measures can be obtained using an appropriate singular value decomposition of the adjacency matrix (Giudici and Spelta, 2016). Centrality measures answer the question of which is the most important unit in a network. They quantify the intuitive notion of nodes' importance in a particular network.

Here, without loss of generality, we consider only the two measures that have a clear and intuitive interpretation, which are the degree centrality and the closeness centrality.

Let $G = (V, E)$ be a graph, with N a set of nodes and E, a set of edges between them, such that any two nodes, say x and y, can be connected by an edge: $e_{xy} = 1$ or not: $e_{xy} = 0$. The degree centrality of a node x is then defined by:

$$d_x = \sum_{y \neq x} e_{xy}.$$

The closeness centrality of a node, instead, is the sum of the length of the shortest paths between the node and all other nodes in the graph. Mathematically:

$$c_x = \frac{1}{\sum_{y \neq x} d_{xy}},$$

where d_{xy} is the distance between nodes x and y.

Giudici and Hadji-Misheva (2017) argue that the above centrality measures could contain useful information that can improve our understanding of loan default determinants. Taking the correlation coefficient as a weight of an edge linking two companies participating in a P2P platforms, the degree centrality would indicate the total number of nodes with which a node is correlated relative to the total it could possibly be connected to. This information could provide important insights as the existence of many significant correlations between companies could capture the presence of joint, underwise unobservable factors. Moreover, the existence of strong, statistically significant correlations linking an active company would several defaulted or bad-performing companies is something that should be included in the credit scoring model.

Giudici and Hadji-Misheva (2017) propose to embed centrality measures into an econometric specification that allows them to become predictive indicators. To this aim, they extend Chinazzi et al. (2012), who incorporate network measures in a linear regression model, to the logistic regression context. This leads to a network-based scoring model, which takes the following form:

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \alpha + \sum_j \beta_j x_{ij} + \gamma g_i,$$

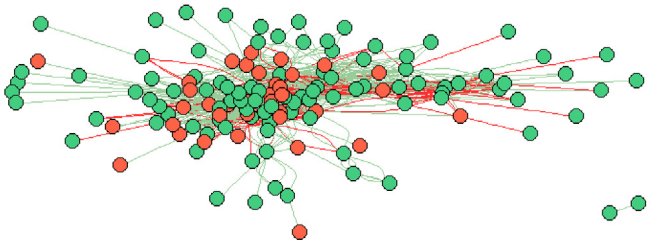

Fig. 1. Correlation network based on the activity indicator.

where p_i is the probability of default, for borrower i, $x_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{ij})$ is a vector of borrower-specific explanatory variables, g_i is the network centrality measure for borrower i and the intercept parameter α and the regression coefficients γ and β_j, for $j = 1, \ldots, J$, are to be estimated from the available data.

We expect that by augmenting the econometric model first by means of centrality measurements, the predictive performance of the scoring model will improve.

3. Application

In this Section provide an example of how correlation network models can improve fintech activity and, specifically, credit risk prediction. The example is taken from Giudici and Hadji-Misheva (2017). Data is collected from an European Fintech specialising in financial consultancy and evaluation of companies' creditworthiness. Although the company does not operate as a P2P platform, it provides the scoring service for investors which in turn is the subject of interest of this study. Specifically, the analysis relies on data on 727 European-based SMEs covering a period of nine years [2007–2015]. The proportion of defaults in the sample is equal to 23%.

As suggested previously, correlation-based network models allow us to observe key properties of the participants in the P2P platform which could be crucial to our ability to identify sharing characteristics and better evaluate credit-risk exposure. In view of the available data, three adjacency matrices have been derived from the correlation matrices using the following variables over period of 9 years: (i) activity, (ii) solvency ratio and (iii) return on equity. In the financial literature, these three ratios are among the most important financial indicators as to the sustainability, liquidity ratio and overall performance of companies. The three 727 $\times$ 727 adjacency matrices (W_1; W_2 and W_3) with elements w_{ij} are obtained on the basis of the correlation matrices for the three time-varying financial indicators (activity, solvency ratio and return to equity) for the set of 727 SMEs included in the sample. Similar as in other studies (Giudici and Spelta, 2016), instead of using a fully connected network, we consider a statistical network in which the edges or links that connect two companies are present on the basis of a partial correlation test that informs whether the corresponding correlation is statistically significant at a given significance level of α.

Fig. 1 shows the network obtained by the correlation matrix using the activity indicator and taking a significance level of α = 0.01. Note that, in the figure, nodes colouring is based on status (red = defaulted companies; green = active companies); edge colouring is based on the correlation sign (green = positive correlation; red = negative correlation). For the sake of visualisation, we only consider correlations greater than 0.90 in absolute value.

Fig. 1 clearly indicates that the network is not fully connected. More importantly, the figure suggest that there exist both positive and negative statistically significant correlations between the activity indicators among companies included in the sample. The existence of many strong statistically significant correlations between companies provided evidence of the existence of joint unobservable forces. The high statistical significance correlation between an active and defaulted company could indicate that they share the same buyers or operate in complementary industries or share other business relationships which are unobservable otherwise. In particular, the fact that an active company is strongly correlated with a company that has defaulted should be relevant for its credit scoring. From the figure we can clearly see that companies which have defaulted (red nodes) are among the most central nodes in the network. Concerning the links, we can interpret them as potentially capturing the effects of competition among the companies included in the sample. Negative correlation could indicate that two companies operate in competition whereas positive correlation could indicate complementarity.

Giudici and Hadji-Misheva go beyond network description and show that, inserting network centrality parameters into a logistic regression model can improve predictive performance. Their results indicate an increase in predictive power, as the AUC value reaches 0.82, from a baseline value 0f 0.69. We believe that this is a clear indicator that network information is crucial in determining borrowers' creditworthiness.

4. Discussion and future research

The paper provides an illustration of what financial data science is, in the wake of the on-going financial technology developments. It shows how data science models and, specifically, correlation network models, can improve credit risk

measurement and, therefore, consumer protection, in peer-to-peer lending, one of the most important fintech business models.

The paper could be extended in several directions. We underline three of them.

A first development is that, while network models identify transmission channels, thus describing how financial contagion spreads through the whole system, time-dependent models associate a specific risk measure to individual institutions, with the aim of predicting what will happen to them in the nearby future, in an early-warning perspective. An important research direction concerns the developments of models that take both perspectives into account.

Another development is to combine different correlation network models into one representation, in a multilayer network perspective.

Last, an important development would be to construct new network centrality parameters, that take into account the fact that nodes may be marked: for example, they should distinguish solvent from insolvent companies.

A broader research extension concerns the application of network models to measure information risk in Financial technology applications different from marketplace lending, such as Robot advisory, Blockchain and cryptocurrencies. And the application to the measurement of cybersecurity. We are currently investigating these issues.

References

Abedifar, P., Giudici, P., Hashem, S., 2017. Heterogeneous market structure and systemic risk: evidence from dual banking systems. J. Financ. Stab. 33, 96–119.

Acharya, V.V., Engle, R., Richardson, M., 2012. Capital shortfall: a new approach to ranking and regulating systemic risks. Am. Econ. Rev.: Pap. Proc. 3, 59–64.

Adrian, T., Brunnermeier, M.K., 2016. Am. Econ. Rev. 106 (7), 1705–1741.

Battiston, S., Delli, G.D., Gallegati, M., Greenwald, B., Stiglitz, J.E., 2012. Liasons dangereuses: Increasing connectivity risk sharing, and systemic risk. J. Econ. Dyn. Control 36 (8), 1121–1141.

Billio, M., Getmansky, M., Lo, A.W., Pelizzon, L., 2011. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. J. Financ. Econ. 104, 535–539.

Brooks, S., Giudici, P., Roberts, G., 2003. Efficient construction of reversible jump MCMC proposal distributions. J. R. Stat. Soc. Ser. B Stat. Methodol. 1, 3–39.

Brownlees, C., Engle, R., 2012. Volatility, Correlation and Tails for Systemic Risk Measurement. Technical Report. New York University.

Calabrese, R., Giudici, P.S., 2015. Estimating bank default with generalized extreme value regression models. J. Oper. Res. Soc. 66 (11), 1783–1792.

Castelo, R., Giudici, P.S., 2001. Improving Markov chain Monte Carlo model search for data mining. Mach. Learn. 50 (1–2), 127–158.

Cerchiello, P., Giudici, P., Nicola, G., 2017. Twitter data models for bank risk contagion. Neurocomputing 264, 50–56.

Chinazzi, M., Fagiolo, G., Reyes, J.A., Schiavo, S., 2012. Post-Mortem Examination of International Financial Network. University of Trento, Italy.

Diebold, F.X., Yilmaz, K., 2014. On the network topology of variance decompositions: Measuring the connectedness of nancial rms. J. Econometrics 182, 119–134.

Giudici, P., Green, P., 1999. Decomposable graphical Gaussian model determination. Biometrika 86 (4), 785–801.

Giudici, P., Hadji-Misheva, B., 2017. Network scoring models for P2P lending. Submitted technical report.

Giudici, P., Parisi, L., 2017. Sovereign risk in the Euro area: a multivariate stochastic process approach. Quant. Finance 17 (12), 1995–2008.

Giudici, P., Sarlin, P., Spelta, A., 2017. The interconnected nature of financial systems: direct and common exposures. J. Bank. Finance (in press).

Giudici, P., Spelta, A., 2016. Graphical network models for international financial flows. J. Bus. Econom. Statist. 34 (1), 126–138.

Giudici, P., Stanghellini, 1999. Bayesian inference for facror analysis models. Psychometrika 66 (4), 577–591.

Guo, Y., Zhou, W., Luo, C., Liu, C., Xiong, H., 2016. Instance-based credit risk assessment for investment decisions in P2P lending. European J. Oper. Res. 246, 417–426.

Lorenz, J., Battiston, S., Schweitzer, F., 2009. Systemic risk in a unifying framework for cascading processes on networks. Eur. Phys. J. B 71 (4), 441–460.

Madigan, D., Raftery, A., 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. J. Am. Stat. Assoc. Financ. Econ. 89 (428), 1535–1546.

Giudici, P., 2003. Applied Data Mining. Wiley, London.

Lauritzen, S., 1996. Graphical Models. Oxford University Press, Oxford.