A prescriptive framework for selecting optimal machine learning algorithms for enterprise-level problems

Prithvi Bhattacharya

University of Wollongong - Dubai Campus, Dubai, UAE

prithvi@uow.edu.au

Abstract. We are living in a world where data is touted to be the new oil and analytics is the combustion engine. Analytics or data science comprises a number of tools and techniques from the fields of statistics, machine learning and broader AI. While there is no dearth of such tools and techniques ranging from random forests to neural networks, problems most enterprises face is that of identify the most suitable tool (s) for a business problem at hand. This paper aims to address this problem by developing a novel, comprehensive framework that identifies the optimal data science tools given the 'nature' and 'type' of the business problem and the constraints on the underlying data used. This framework can be an effective device for enterprises to select the suitable data science tools and techniques to apply to their problems. The intention is to empirically test this framework in future research and develop a Natural language Processing (NLP) implementation of this prescriptive framework.

Keywords: machine learning, data science, analytics.

1. Introduction

The current decade has witnessed the emergence of and widespread attention to a term 'Data Science', sometimes also known as, 'analytics' [9,18,33]. This term is not ground-breaking; it is an amalgamation of classical disciplines like Artificial Intelligence (including machine learning) statistics, data mining and even operations research [42]. However, the ground-breaking aspect of this term is that it has made these above disciplines mainstream by allowing application of the certain tools and techniques of these disciplines to solve a wide range of business problems [13]. It has also been found that data science shows an enormous promise to contribute to both the top-line and the bottom-line growth of businesses [23,34].

It may be noted that the above examples are from a miniscule proportion of the total number of organizations in the marketplace; many large and most medium-sized organizations are yet to tap into the potential of data science for solving their problems. There are a number of issues that impede businesses to embrace and harness the potential of AI technologies and data science like the lack of understanding of the plethora of models and tools provided, privacy concerns, ethics concerns, economic deterrents, legal limitations and many others [5]. Discussing all the challenges of data science is outside the scope of this study.

However, a particular glaring issue that would affect any organization anywhere in the world willing to use data science is: the lack of understanding of the plethora of tools provided to choose the right one for the problem at hand [24,43]. As mentioned earlier, data science is a consolidation of a wide range of tools and techniques (most often in the form of models and algorithms) from artificial intelligence, data mining, and statistics. More precisely, data science incorporates a wide range of different techniques, models and algorithms. From Support Vector Machines, k-Nearest Neighbour, Random Forests, K-Means, Spectral Clustering and many others from the field of computer science to Linear/Logistic Regression, Exponential Smoothing, Probabilistic Topic Modeling (like LDA) from the field of statistics; the list is a very extensive one [24]. Many of these tools have been shown to solve certain specific types of problems well (in terms of accuracy and/or efficiency) under a certain specific set of conditions; these are not intended to be applied indiscriminately. Consequently, it is crucial to understand which of these tools and techniques can be applied to which business problem, in order to solve it successfully. Even using an exceptionally well performing data science tool for the wrong purpose will be futile at best and detrimental at worst for an organization. It may be noted here that the focus of this study is to solve problems typically faced by business organizations, not so much the ones faced by other types of entities or individuals. To this end, the purpose of this study is to answer the research question below.

How can we identify the most suitable modelling technique(s) from a large variety of data science tools to address specific enterprise problems at hand?

The remainder of this paper outlines a review of the relevant literature and proposes a framework that attempts to answer the research question. Subsequently the paper entails the contribution and an outline of the forthcoming courses of action and implications for future research.

2. Reviewing the literature

Conducting a literature review on data science reveals that given the term 'data science ' is relatively new, not much is found in the literature with a search for that term. However, on conducting a broader search on terms that are similar and convey the essence of 'data science', i.e. data mining, analytics, machine learning, statistical modeling and similar other terms, a considerable number of studies on the tools, techniques and models from the fields of data mining, statistics and machine learning were found. As a matter of fact, the vast number of studies are well summarized in a number of systematic literature reviews. For example, there is a wealth of literature with systematic reviews done on different data mining techniques on different domain ranging from healthcare to education to intrusion detection [7,15,17,28,36,39]. Similarly, a number of useful systematic literature reviews are found on the tools and techniques available from the field of statistics [10,20,27,32]. Further, in the area of machine learning, there are a number of studies comparing a number of machine learning tools and models in a specific scenario each, ranging from phishing detection to bioinformatics to customer churn to digital soil mapping [1,21,40,41].

Given the plethora of tools, techniques and models at one's disposal, it is of utmost importance to identify the right 'horse for the course', or, the right tool for the problem at hand. While the literature is replete with studies on different tools, techniques and models used in the different disciplines and their application, every study tends to focus on one (or a few tools) like k-means clustering, support vector machine, regression and others, and their application in a specific domain like fraud detection, genomics or manufacturing. So it is extremely difficult to have a cross-sectional understanding of these tools to solve different problems. In other words, the majority of studies in the literature are depth-based. However, the primes of the paper is not to dig deep into one particular tool or an application domain, but instead to explore the wide variety of tools, techniques and models and their applicability to different types of business problems: essentially a breadth-based approach. So a search for breadth-based studies was necessary to be conducted.

A further detailed search revealed a small number of papers that are indeed broad-based, in the sense they discuss a wide variety of tools, techniques and models from multiple disciplines [25,26]. One such key study is one that depicts the classification of the methods for predictive questions in three categories: Probabilistic Models, Machine Learning/Data Mining and Statistical Analysis and the classification of the methods for prescriptive questions in six categories: Probabilistic Models, Machine Learning/Data Mining, Mathematical Programming, Evolutionary Computation, Simulation and Logic-based Models [25]. However, the limitation of this study is that does not refer to a wide variety of different types of problems (like classification, estimation, clustering and others) that fall under either predictive or prescriptive questions; a generic umbrella approach is used which limits its applicability in different types of problems.

A second study reports on a number of different tools and models used for a specific type of descriptive and predictive question: clustering [6]. This provides a comprehensive list of the different tools available but with the limitation that they all belong to a single type of analytics question: clustering. A similar approach is taken by another study to analyse another problem, image detection, using different tools and models. Its limitation is also that it restricts itself to a specific kind of problem [44]. A yet another similar study has explored the use of different models and techniques, but again restricted to one domain: the financial services [3]. One more study reports similar content, but in the area of supply chain management [30].

Yet another study contains an in-depth overview of different analytics, algorithms and platforms. However, the study does not delve into the application of these tools and models to solve business problems [22]. This limitation is also applicable to a number of other studies that limit their domains to scientific computing problems as opposed to business organizational ones.

So overall, it is evident that there is a dearth of studies that explore a cross-section of data science tools, techniques and models (across various fields) vis –a vis their applicability in answering different types of questions (namely descriptive, predictive and prescriptive) as well as different sub-types of questions (namely classification, estimation, clustering etc.). Also, a lack of focus on solving business organizational problems by comparing a number of tools and models was found in the available literature.

3. Towards a framework to use data science to find optimal solutions to business problems.

This section describes a framework that can be used to identify optimal tools and models from data science for solving business problems. Data science equips enterprises with platforms for deconstructing the problem into smaller problems, and the tools and techniques for addressing them. As with the field of data science itself, these tools and models are sourced from the varied spheres of statistics, computer science, data mining, operations research among others. The purpose of this framework is provide a holistic framework to help make decisions about the choice of tools and models to apply to a given problem at hand.

The building blocks of this framework are:

a) well-formed business problem at hand
b) access to potentially relevant data to solve it
c) a set of tools, techniques and models to solve the business problem

The first building block is the business problem. The problem can be analysed using an industry standard specification called the Business Motivation Model (BMM) from Object management Group that provides a hierarchy of the goals of an organization/entity, decomposed into several levels [31].

According to the BMM Specification, a ‘Vision’ is the ultimate, possibly unattainable, state the enterprise would like to achieve For example: “To become the first choice vehicle rental company for customers in Asia”. A vision will be implemented through a number of goals. A ‘Goal’ is a statement about a state or condition of the organization to be achieved and/or maintained via suitable approaches. E.g. “To become a ‘luxury brand’ vehicle rental company, operating in the same realm as other premier companies such as Avis and Apex. A goal should be neither too broad nor too narrow; it should be suitable to be quantified by Objectives. An ‘Objective’ is a statement of a specific, attainable, realistic, time-targeted, and measurable target that the enterprise will have to meet for achieving its goals. The objective must be framed meet the widely accepted SMART criterion: specific, measurable, achievable, relevant, and timed, as initially suggested by Peter Drucker in his seminal book, Practice of Management [14]. This has evolved over time and is considered the gold standard in setting objectives. E.g. Within six months, 10% increase in product sales, or during 4th quarter of current year, no more than 1% of rentals need the car to be replaced because of mechanical breakdown [31]. For each such objective, there can be one or more underlying ‘Contextual Questions’ that need to be answered to be able to attain the objective. These questions, like their parent objective, have to meet the SMART criteria.

The second building block is the access to potentially relevant data that can be used to answer the question. The source of such data must be credible and verifiable. Such data must be adequately sourced, cleansed, pre-processed and reviewed for use [33].

The third building block is a set of tools, techniques and models from data science for answering the contextual questions above. It may be noted that this framework does not claim to be list an exhaustive set of all possible models and tools available to solve business problems. Instead, the intention of the framework is to provide a comprehensive subset of the key list of tools and models that have been found to have been applied on business problems as evidenced in the literature [25,33]. Such a list, though not exhaustive, can be a useful suggestive resource for applying the most suitable models for a given problem from a plethora of available solutions The framework is intended to identify and recommend tools, techniques and models from data science for answering the contextual questions to eventually address the objectives goals and vision of the organization.

3.1. Contextual questions: nature and type

The proposed framework consists of a matrix that is tiered, with a tier each referring to Descriptive, Predictive and Prescriptive contextual question, and are named as such. The first task in identifying an optimal model is to identify the nature of the contextual question: Descriptive (sometimes called Exploratory or Diagnostic), Predictive or Prescriptive [25]. Typically, a ‘Descriptive’ question seeks answers that describe the data and any possible association between the data. It may be noted that there is an extension to descriptive analytics named “diagnostic analytics” which is related to the question “Why did it happen?” [25]. ‘Predictive’ questions, go beyond that and seek answers about predicting results about the future based on existing data. Finally, ‘Prescriptive’ questions aim to find optimal results for a given problem. These aim to answer, what should happen, not just what would happen [4,11]. This task of categorizing a question based on its nature as Descriptive, Predictive or Prescriptive is non-trivial and is usually performed by humans with expert knowledge. Each of these tiers (Descriptive, Predictive and Prescriptive) have a number of different sub-types of questions underneath it, with a set of tools, techniques and models associated with it. These are listed as rows in under each of the tiers.

While there are a number of modeling techniques in solving different descriptive questions, all such questions fall broadly under a finite set of types [33]. The task of determining the type of the question under each tier is non-trivial and is usually performed by humans with expert knowledge.

3.2. Underlying data: constraints

Also, it may be noted that each of these tools, techniques and models are bound by certain assumptions about the underlying data. These include the type of data (or response) being studied (numeric or categorical), the underlying probability distribution that has a fixed set of parameters, multi-collinearity, homoscedasticity, outlier, distribution of residuals, sample sizes and many others [16]. These assumptions, also called constraints in this context, are listed on the top of the matrix as columns. Wherever a tool is subject to a constraint, it is marked in the appropriate columns. So, a tool with no marks across the columns in the most generally applicable with almost no constraints they involve few (if any) assumptions of structure and distribution. In contrast, a tool with many marks across the columns are heavily constrained by assumptions; these are usually the so-called non-parametric or semi-parametric models.

It may be noted that the most common, relevant assumptions/constraints about the models, as identified in the literature, are listed as the columns of the matrix; it may not be an exhaustive list of all assumptions about all variants of a model. This simplification is necessary in the interest of the clarity and presentability of the matrix for any practical purposes.

Once the nature and the type of the question is decided, and the constraints of the underlying data are identified, the proposed framework can be used to identify the suitable tool(s) or model(s) to answer the data science question at hand, by referencing the appropriate row in the appropriate tier. These tiers are described below.

Tier A: Descriptive Question

The answers to these questions usually take the form of describing the data and are often used for exploring data for further analysis. These typically are questions that are of the type “What has happened or is happening?” and “Why does it happen?” [25,26]. Such questions are mostly well answered by the tools mentioned in the Tier A of the table. Tier A of the framework enumerates the key different types of descriptive questions and then attempts to suggest the feasible modeling solution(s) for each of those types [35]. These tools are subject to the constraints, marked for each tool, listed in the columns of the table. Out of those feasible models suggested by the framework, a model can be selected as the optimal one based on experimentation and context knowledge. A brief description each of these different types of questions is given below.

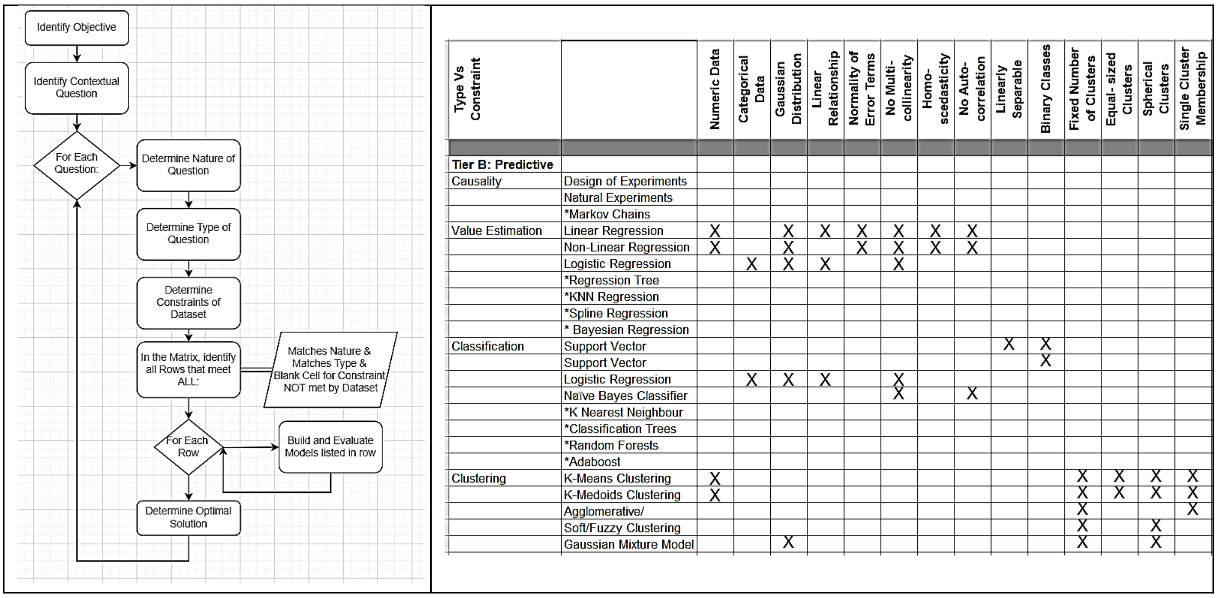
Tier B: Predictive Question

The answers to these questions usually take the form of predicting responses for future instances. These typically are questions that are of the type “What will happen?” and “Why will it happen?” in the future [25,26,29]. This tier of the framework has an embedded, built-in ‘inference’ component (with tools like confidence intervals and hypothesis testing) that applies across all models of this type. This is to make the models applicable to the population even if they are built using data from samples.

While there are a number of modeling techniques in solving different questions, all such questions fall broadly under a finite set of categories or types [33]. Tier B of the framework enumerates those different types of predictive questions and then attempts to suggest the feasible modeling solution(s) for each of those types [25,29,33]. These tools are subject to the constraints, marked for each tool, listed in the columns of the table. Out of those feasible models suggested by the framework, a model can be selected as the optimal one based on experimentation and context knowledge. A brief description each of these different types of questions is given below.

Tier C: Prescriptive Question

The answers to these questions usually take the form of suggesting a feasible solution for a given situation. These typically are questions like “What should I do?” and “Why should I do it? [19,25]. While there are a number of modeling techniques in solving different questions, all such questions fall broadly under a finite set of categories or types [25]. Tier C of the framework enumerates those different types of questions and then attempts to suggest the feasible modeling solution(s) for each of those types [8,12,19,35,38]. These tools are subject to the constraints, marked for each tool, listed in the columns of the table. Out of those feasible models suggested by the framework, a model can be selected as the optimal one based on experimentation and context knowledge. In summary, the recommended steps of using the proposed framework to identify an optimal solution to business problems is given below:


Identify Objective
Identify Contextual Question
For Each Question:
Determine Nature of Question
Determine Type of Question
Determine Constraints of Dataset
In the Matrix, identify all Rows that meet ALL:
Matches Nature & Matches Type & Blank Cell for Constraint NOT met by Dataset
For Each Row
Build and Evaluate Models listed in row
Determine Optimal Solution

Type Vs Constraint		Numeric Data	Categorical Data	Gaussian Distribution	Linear Relationship	Normality of Error Terms	No Multi-collinearity	Homo-scedasticity	No Auto-correlation	Linearly Separable	Binary Classes	Fixed Number of Clusters	Equal- sized Clusters	Spherical Clusters	Single Cluster Membership
Tier B: Predictive															
Causality	Design of Experiments														
	Natural Experiments														
	*Markov Chains														
Value Estimation	Linear Regression	X		X	X	X	X	X	X						
	Non-Linear Regression	X		X		X	X	X	X						
	Logistic Regression		X	X	X		X								
	*Regression Tree														
	*KNN Regression														
	*Spline Regression														
	* Bayesian Regression														
Classification	Support Vector									X	X				
	Support Vector										X				
	Logistic Regression		X	X	X		X								
	Naïve Bayes Classifier						X		X						
	*K Nearest Neighbour														
	*Classification Trees														
	*Random Forests														
	*Adaboost														
Clustering	K-Means Clustering	X										X	X	X	X
	K-Medoids Clustering	X										X	X	X	X
	Agglomerative/											X			X
	Soft/Fuzzy Clustering											X		X	
	Gaussian Mixture Model			X								X		X	

Figure 1. A framework for selecting the optimal modelling technique for enterprise-level problems.

4. Conclusion and implications for future

This paper explores a contemporary topic of using artificial intelligence techniques, data science in particular, to solve business problems and to this end, outlines a review of the relevant literature and proposes a new framework. The contribution of this study is that it attempts a holistic, comprehensive framework that attempts to guide the selection of suitable data science tools, techniques and models to solve a variety of business problems. The framework explains how the different data science models can help answer 'contextual' questions of different nature (descriptive, predictive and prescriptive) and of different types (classification, estimation, clustering and many others). Answering these questions will help organizations meet their objectives and in turn their goals and vision. Thus this study provides practitioners a practical guide to identify and implement the specific data science solutions (from a plethora of widely advertised tools, techniques and models) that would really assist to solve their specific business problems.

References

[1] Abu-Nimeh, S., Nappa, D., Wang, X., Nair, S., 2007. A comparison of machine learning techniques for phishing detection, in: Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit. pp. 60–69.

[2] Aminikhanghahi, S., Cook, D.J., 2017. A Survey of Methods for Time Series Change Point Detection. Knowl Inf Syst 51, 339–367. https://doi.org/10.1007/s10115-016-0987-z

[3] Andriosopoulos, D., Doumpos, M., Pardalos, P.M., Zopounidis, C., 2019. Computational approaches and data analytics in financial services: A literature review. Journal of the Operational Research Society 70, 1581–1599.

[4] Appelbaum, D., Kogan, A., Vasarhelyi, M., Yan, Z., 2017. Impact of business analytics and enterprise systems on managerial accounting. International Journal of Accounting Information Systems 25, 29–44.

[5] Berendt, B., 2019. AI for the common good?! pitfalls, challenges, and ethics pen-testing. Paladyn, Journal of Behavioral Robotics 10, 44–65.

[6] Berkhin, P., 2006. A survey of clustering data mining techniques, in: Grouping Multidimensional Data. Springer, pp. 25–71.

[7] Bertoni, A., Larsson, T., 2017. Data mining in product service systems design: Literature review and research questions. Procedia CIRP 64, 306–311.

[8] Bertsimas, D., Kallus, N., 2020. From predictive to prescriptive analytics. Management Science 66, 1025–1044.

[9] Brynjolfsson, E., Mcafee, A., 2017. The business of artificial intelligence. Harvard Business Review 1–20.

[10] Cranmer, S.J., Leifeld, P., McClurg, S.D., Rolfe, M., 2017. Navigating the range of statistical tools for inferential network analysis. American Journal of Political Science 61, 237–251.

[11] Davenport, T.H., 2015. The rise of automated analytics. The Wall Street Journal.

[12] Dey, S., Gupta, N., Pathak, S., Kela, D.H., Datta, S., 2019. Data-Driven Design Optimization for Industrial Products, in: Datta, S., Davim, J.P. (Eds.), Optimization in Industry: Present Practices and Future Scopes, Management and Industrial Engineering. Springer International Publishing, Cham, pp. 253–267.

[13] Dhar, V., 2013. Data science and prediction. Communications of the ACM 56, 64–73.

[14] Drucker, P.F., 1954. Practice of Management.

[15] Dutt, A., Ismail, M.A., Herawan, T., 2017. A systematic review on educational data mining. Ieee Access 5, 15991–16005.

[16] Gao, C., Sun, H., Wang, T., Tang, M., Bohnen, N.I., Müller, M.L., Herman, T., Giladi, N., Kalinin, A., Spino, C., 2018. Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease. Scientific reports 8, 1–21.

[17] Genc-Nayebi, N., Abran, A., 2017. A systematic literature review: Opinion mining studies from mobile app store user reviews. Journal of Systems and Software 125, 207–219.

[18] Ghosh, B., Durg, K., Deo, A., Fernandes, M., 2018. Want the Best Results From AI? Ask a Human [WWW Document]. URL https://sloanreview.mit.edu/article/want-the-best-results-from-ai-ask-a-human/ (accessed 3.2.20).

[19] Gröger, C., Schwarz, H., Mitschang, B., 2014. Prescriptive Analytics for Recommendation-Based Business Process Optimization, in: Abramowicz, W., Kokkinaki, A. (Eds.), Business Information Systems, Lecture Notes in Business Information Processing. Springer International Publishing, pp. 25–37.

[20] Guruputranavar, N., 2019. An investigative study on the application of different statistical tools and methods for optimizing the hole machining on polymer matrix composites–A review. International Journal of Mechanical Engineering and Technology 10, 1033–1043.

[21] Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77.

[22] Kejariwal, A., Kulkarni, S., Ramasamy, K., 2017. Real Time Analytics: Algorithms and Systems. arXiv:1708.02621 [cs].

[23] Kolbjørnsrud, V., Amico, R., Thomas, R.J., 2016. How artificial intelligence will redefine management. Harvard Business Review 2, 1–6.

[24] Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence 5, 221–232.

[25] Lepenioti, K., Bousdekis, A., Apostolou, D., Mentzas, G., 2020. Prescriptive analytics: Literature review and research challenges. International Journal of Information Management 50, 57–70.

[26] Lu, J., Chen, W., Ma, Y., Ke, J., Li, Z., Zhang, F., Maciejewski, R., 2017. Recent progress and trends in predictive visual analytics. Front. Comput. Sci. 11, 192–207

[27] Maleki, M.R., Amiri, A., Castagliola, P., 2017. Measurement errors in statistical process monitoring: a literature review. Computers & Industrial Engineering 103, 316–329.

[28] Malik, M.M., Abdallah, S., Ala'raj, M., 2018. Data mining and predictive analytics applications for delivery of healthcare services: a systematic literature review. Annals of Operations Research 270.

[29] Mishra, N., Silakari, S., 2012. Predictive analytics: A survey, trends, applications, oppurtunities & challenges. International Journal of Computer Science and Information Technologies 3, 4434–4438.

[30] Nguyen, T., Zhou, L., Spiegler, V., Ieromonachou, P., Lin, Y., 2018. Big data analytics in supply chain management: A state-of-the-art literature review. Computers & Operations Research 98, 254–264. https://doi.org/10.1016/j.cor.2017.07.004

[31] Object Management Group, 2015. Business Motivation Model Version 1.3 [WWW Document]. URL https://www.omg.org/spec/BMM/1.3/PDF (accessed 5.24.20).

[32] Peres, F.A.P., Fogliatto, F.S., 2018. Variable selection methods in multivariate statistical process control: A systematic literature review. Computers & Industrial Engineering 115, 603–619.

[33] Provost, F., Fawcett, T., 2013. Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc.

[34] Ransbotham, S., Kiron, D., Gerbert, P., Reeves, M., 2017. Reshaping business with artificial intelligence: Closing the gap between ambition and action. MIT Sloan Management Review 59.

[35] Ratner, B., 2017. Statistical and Machine-Learning Data Mining:: Techniques for Better Predictive Modeling and Analysis of Big Data. CRC Press.

[36] Salo, F., Injadat, M., Nassif, A.B., Shami, A., Essex, A., 2018. Data mining techniques in intrusion detection systems: A systematic literature review. IEEE Access 6, 56046–56058.

[37] Scherer, M.U., 2015. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. Harv. JL & Tech. 29, 353.

[38] Shroff, G., Agarwal, P., Singh, K., Kazmi, A.H., Shah, S., Sardeshmukh, A., 2014. Prescriptive information fusion, in: 17th International Conference on Information Fusion (FUSION). Presented at the 17th International Conference on Information Fusion (FUSION), pp. 1–8.

[39] Silva, C., Fonseca, J., 2017. Educational Data Mining: a literature review, in: Europe and MENA Cooperation Advances in Information and Communication Technologies. Springer, pp. 87–94.

[40] Tan, A.C., Gilbert, D., 2003. An empirical comparison of supervised machine learning techniques in bioinformatics, in: Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003-Volume 19. Australian Computer Society, Inc., pp. 219–222.

[41] Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., Chatzisavvas, K.C., 2015. A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory

[42] Van Der Aalst, W., 2016. Data science in action, in: Process Mining. Springer, pp. 3–23.

[43] Waller, M.A., Fawcett, S.E., 2013. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. Journal of Business Logistics 34, 77–84.

[44] Xiao, H., Rasul, K., Vollgraf, R., 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv:1708.07747 [cs, stat].

[45] Zhang, M., Chen, Y., 2018. Link Prediction Based on Graph Neural Networks, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 31. Curran Associates, Inc., pp. 5165–5175.