Business & Information Systems Engineering manuscript No.
(will be inserted by the editor)

Closing the Gap between Experts and Novices using Analytics-as-a-Service

An Experimental Study

Jasmien Lismont · Tine Van Calster · María Óskarsdóttir · Seppe vanden Broucke · Bart Baesens · Wilfried Lemahieu · Jan Vanthienen

Final submitted draft - February, 2018

Abstract Generating insights and value from data has become an important asset for organizations. At the same time, the need for experts in analytics is increasing and the number of analytics applications is growing. Recently, a new trend has emerged, i.e. analytics-as-a-service platforms, that makes it easier to apply analytics both for novice and expert users. In this study, we approach these new services by conducting a full-factorial experiment where both inexperienced and experienced users take on an analytics task with an analytics-as-a-service technology. Our research proves that although experts in analytics still significantly outperform novices, these web-based platforms do offer an advantage to inexperienced users. Furthermore, we find that analytics-as-a-service does not offer the same benefits across different analytics tasks. That is, we observe better performance for supervised analytics tasks. Moreover, this study indicates that there are significant differences between novices. The most important distinction lies in the approach they take on the task. Novices who follow a more complex, although structured, workflow behave more similarly to experts and, thus, also perform better. Our findings can aid managers in their hiring and training strategy with regards to both business users and data scientists. Moreover, it can guide managers in the development of an enterprise-wide analytics culture. Finally, our results can inform vendors about the design and development of these platforms.

Keywords Analytics-as-a-service · Automated analytics · Data analytics · Experimental study · Novices

J. Lismont · T. Van Calster · M. Óskarsdóttir · S. vanden Broucke · B. Baesens · W. Lemahieu · J. Vanthienen
KU Leuven, Dept. of Decision Sciences and Information Management,
Naamsestraat 69, B-3000 Leuven, Belgium
Tel.: +32 16 326 8 84
E-mail: Bart.Baesens@kuleuven.be

B. Baesens
University of Southampton, United Kingdom

1 Introduction

Data analytics, where advanced techniques are applied to data in order to gain novel insights, has become an important asset in companies for achieving competitive advantage (Baesens, 2014; Davenport & Harris, 2007). Recently, it has even become a necessary capability in order to stay competitive in the market (Ransbotham et al., 2016). This leads to the necessity of growing increasingly larger teams of specialized analysts (Lismont et al., 2017), i.e. data scientists, causing increasing concerns that the necessary skills are scarcely available in the market (Chen et al., 2012; Zorrilla & García-Saiz, 2013). At the same time, two trends have been developing which offer a potential solution. Firstly, there is a current tendency of empowering business experts who are nevertheless novices when it comes to analytics (Alpar & Schulz, 2016), and, similarly, of making analytics more accessible (Gartner, 2015). Debortoli et al. (2014) emphasize again that business knowledge is equally important as technical skills. Accessible analytics platforms allow companies to leverage business expertise and can at the same time provide an answer to the predicted shortages of analytics experts (Leavitt, 2013; Zorrilla & García-Saiz, 2013). In this context, Alpar & Schulz (2016) mention the development of new web-based applications, i.e. analytics-as-a-service (AaaS) platforms. This leads us to a second trend, namely that of (partially) automating analytics. Most AaaS platforms include services that offer an efficient, data-driven and cloud-based solution to business problems ranging from data storage and preparation, to model deployment and evaluation.

This paper aims to investigate whether data analytics can in fact successfully be made more accessible to a broader audience by means of semi-automated analytics. For this purpose, an experimental design is set up where experts and novices in analytics undertake an analytics task by means of AaaS. Firstly, we assess how well novices perform when applying AaaS for an analytics task compared to a random baseline model. This will allow us to research whether novices can actually achieve acceptable performance. These results are also contrasted with the results that the experts achieved when using the same platform. Secondly, the paper investigates whether certain tasks are more approachable with AaaS by taking both problem setting and data quality into account. Finally, as AaaS is suggested as usable by business users, the performance of the novice users is further analyzed by measuring the influence of user characteristics, task characteristics and the user's approach to the task on model accuracy.

The results of these analyses lead to three main contributions. (1) Our findings illustrate that while experts still significantly outperform novices with regards to an analytics task, the application of AaaS platforms allows novices to perform significantly better than a random baseline model. Although it might be expected to perform better than a random model, simply completing an analytics project is not straightforward for amateurs. Moreover, if business users can achieve decent performance, this can contribute to a data culture and to closer collaborations with analytics experts. (2) Nevertheless, this research also illustrates that supervised tasks are more approachable in the context of AaaS platforms, regardless of the level of expertise of the user. (3) Finally, the performance of novice users is influenced by both user and task characteristics, but is mostly defined by the user's task approach. The task approach of the best performing novice users is

complex and more similar to the approach of expert users. These findings can be used to guide managers in trainings, but also to inform AaaS vendors.

The following section covers related research. Consecutively, in Section 3, the methodology with the experimental design and set-up is discussed, as well as the applied techniques. Section 4 presents the results together with a discussion of their implications and validation.

2 Related research

Our paper focuses on platforms which improve user-friendliness of data analytics applications for which specialized statistics and machine learning techniques are applied to generate new insights from data. This new information can be extracted for either existing business problems where the goal is to predict an outcome, e.g. churn prediction or credit scoring; or for problems that try to derive structure and patterns from data sources, e.g. customer segmentation. These business problems are also known as supervised and unsupervised problems, respectively. In this paper, we zoom in on AaaS, which aims to introduce analytics to the masses and enlarge the application domain from analytics experts to (unexperienced) business users or novices. In what follows, we first discuss the definition of AaaS applied in this paper. Next, we cover the advantages and disadvantages of these platforms.

2.1 Analytics-as-a-service defined

AaaS, sometimes called 'agile analytics', has previously been defined as generating insights from data wherever this data may be located and to turning a "general purpose analytical platform into a shared utility" (Demirkan & Delen, 2013). This definition relates AaaS to other concepts such as cloud computing, utility computing and on-demand services. Furthermore, AaaS relates to the concept of self-service business intelligence (BI), or services that allow users to perform their own BI. Weinhardt et al. (2009) observe a current trend in cloud computing of closing the gap between business and technology. Nevertheless, BI is a much wider field than data analytics. As such, Alpar & Schulz (2016) refer to three levels of self-service BI: usage of information, creation of information, and creation of information resources. Each level demands increasing system support and self-reliance of the user. Imhoff & White (2011) executed a survey on the use of self-service BI, in which they discovered three maturity levels: basic BI, standard BI and advanced BI. Only the last level corresponds with the definition of data analytics above.

In this paper, we define AaaS as a cloud-based service designed to support the entire data analytics process from data preparation to interpretation. More specifically, our attention goes to platforms that offer both descriptive and predictive machine learning techniques by means of a web-based portal. These semi-automated analytics platforms offer a user-friendly interface with drag-and-drop modules which automate techniques with the possibility of setting parameters. Additionally, they typically provide numerous templates and extensive documentation to guide users in their analytics projects. This definition of AaaS, however, does not fit nicely within the definition of cloud computing by NIST (Mell & Grance,

2011). One can position AaaS under either software-as-a-service or platform-as-a-service, depending on the characteristics of the service itself. Moreover, vendors frequently offer multiple deployment models, depending on the requirements of the user, such as the option of a private cloud.

2.2 A motivation for analytics-as-a-service

AaaS has some interesting characteristics which make it an attractive alternative for standalone analytics tools. Some of these characteristics are related to the 'cloud' or 'as-a-service' aspect of AaaS. Firstly, AaaS, in general, offers a usage-based pricing model (Armbrust et al., 2010; Chen & Wu, 2013; Demirkan & Delen, 2013). This type of model allows gradual analytics deployment and may even enable the execution of new ideas that were not possible before (Chen & Wu, 2013; Leavitt, 2013). This advantage demonstrates the popularity of on-demand services for start-ups and small- and medium-sized companies (Gupta et al., 2013; Marston et al., 2011; Weinhardt et al., 2009), but it might also deliver opportunities for incumbent firms. Larger organizations also struggle to dedicate the necessary resources for processing data in a timely manner (Demirkan & Delen, 2013). Secondly, AaaS includes fast development and deployment of analytics models (Chen & Wu, 2013; Demirkan & Delen, 2013). The reusability of software components and analytical processes contributes to a more cost efficient application. This also facilitates inter- and intra-enterprise access to proven and shared expertise, since resources, such as data and analytical results, are more easily shared (Chen et al., 2011). Thirdly, capacity constraints are reduced (Chen & Wu, 2013), as pooled resources enable flexible analytics capacities. These resources are easier to maintain and software can be upgraded in a more flexible manner (Elazhary, 2014). Additionally, AaaS offers better scalability (Demirkan & Delen, 2013; Elazhary, 2014; Marston et al., 2011) in comparison to standalone tools. In general, ease of use and convenience are the biggest factor mentioned by smaller companies to adopt cloud services (Gupta et al., 2013). Finally, Leavitt (2013) explicitly mentions as an advantage of AaaS that it will no longer be necessary to have employees with data analytics related skills. This would reduce human capacity constraints and offer an answer to predicted shortages of data scientists (Chen et al., 2012), although this statement is criticized, for example, by Davenport (2014, p.110).

2.3 The challenge of analytics-as-a-service

AaaS also comes with a number of challenges which prevent a straightforward application. Firstly, privacy and security risks are encountered (Armbrust et al., 2010; August et al., 2014; Chen & Wu, 2013; Demirkan & Delen, 2013; Elazhary, 2014; Marston et al., 2011; Weinhardt et al., 2009). Data can be regarded as a unique asset for companies and a leverage for competitive advantage. Companies are worried about how data privacy and security are handled in AaaS (Lismont et al., 2015). Company politics might moreover explicitly prohibit the use of public clouds for confidential data. Secondly, data control is preferred, which leads to the concept of accountability. Legal regulations are currently not following market demand and are country-specific (Demirkan & Delen, 2013; Marston et al., 2011).

Moreover, companies that try to reduce the risk by encrypting their data, might be facing technical challenges (Demirkan & Delen, 2013). In this context, Jaatun et al. (2016) emphasize the importance of educating end-users on responsible data stewardship. Thirdly, companies might be confronted with switching costs (Chen & Wu, 2013). Once data is in the cloud, it is often hard to get it out again, leading to a data lock-in (Armbrust et al., 2010). Companies already have hardware and software in place and thus new implementations need to be able to interact with legacy tools. Moreover, concerns may exist about service availability (Armbrust et al., 2010; Demirkan & Delen, 2013; Marston et al., 2011), as companies who use AaaS, want fast access at all times. Data transfer bottlenecks, when data is uploaded or downloaded from the server, can occur when not enough capacity is available (Armbrust et al., 2010; Marston et al., 2011). Finally, there are concerns with regards to the validity of the analytical insights. If business users apply AaaS, will they still know which data drives their insights? Managers are often reluctant about methods that they cannot fully comprehend (Baesens, 2014). Non-experienced users may not know which techniques the platform employs or how they work, which results in a black box outcome regardless of whether the techniques themselves are black or white box in nature. Upon choosing the right AaaS, a choice might therefore be required between ease of use and comprehensibility of the underlying techniques. In relation to this, previous research has questioned which level of expertise in BI is necessary for these users in order to produce reliable insights. Alpar & Schulz (2016) acknowledge the risk that business users are often not able to clearly formulate their questions nor validate their solutions with regards to analytics. Therefore, we chose to explicitly address the level of analytics expertise in this paper.

3 Methods

In order to analyze the impact of AaaS, we set up an experiment. Firstly, in Section 3.1, we describe the participants and the AaaS technology employed. Participants are analytics novices and experts, who both perform this experiment with a specific AaaS platform. Next, we discuss the design of the experiment in Section 3.2. The design is a full-factorial experiment with three factors, namely expert level, the analytics task and data quality. Consecutively, we discuss how we measure the task performance of the participants. Finally, in Section 3.3, we explain how we extracted information from the experiment and how we analyze these data by means of a factor analysis, linear regression and process analysis.

3.1 Experimental set-up

There are two types of participants in our experiment, namely novices and experts. For each group, the sample size, subject mortality, background, recruitment process and environment are described in the next paragraphs and summarized in Table 1.

Novices are represented by undergraduate, graduate and exchange students at a Belgian university, KU Leuven, and a Belgian college, UCLL, and have no

educational background or work experience concerning analytics. In Belgium, universities deliver academic degrees, while colleges focus on professional degrees. Both schools belong to KU Leuven association[1]. Students were attending a study program in the domain of business economics, statistics or computer science at the time of the experiment. In total, 92 novices participated, of which ten were excluded due to incomplete results. Novices were recruited by means of communication through a selection of relevant courses. Participation was not mandatory, but a reward was offered by means of random draw. The experiment took place at six different timings in a supervised classroom setting. Each group received a maximum of three hours to finish the experiment using a desktop computer running a customized Java application. No communication between participants was allowed. For each student, their knowledge on marketing, finance and statistics was tested by means of five multiple choice questions (MCQTest)[2], leading to a mean score of 2.13. Note that the characteristic 'work experience' for novices does not relate to analytics experience, but to work experience in general.

Experts have at least one year of work experience in analytics. In total, 22 experts participated, of which none dropped out. They were recruited through LinkedIn based on their profile in analytics, and came from different countries and industries. University sponsoring was made apparent to the participants and all experts were offered a reward. Experts were given the opportunity to participate remotely by using their own device. By means of a server connection, they were able to run the same customized application as the novices. Participants were requested to finish the experiment in one go, and within three hours. Although we were not able to enforce this last requirement due to the set-up, almost all experts finished the experiment within three hours with an average of 1 hour and 54 minutes and a median of 1 hour and 41 minutes.

The novices and experts were requested to solve a given business problem using AaaS. In this experiment, only one platform was selected in order to avoid a benchmarking of different vendors, as this is not the goal of this research. Concretely, Azure Machine Learning Studio of Microsoft[3] (Azure) was selected (Van Calster et al., 2016). This platform is easy-to-use by means of drag-and-drop analytics elements (Jaatun et al., 2016). Additionally, it allows for an end-to-end solution, with the possibility to include Python and R code during the construction of the analytics model. It also offers tutorials and documentation on the different techniques and their possible applications. Finally, Azure is popular among data scientists (Lismont et al., 2015; Van Calster et al., 2016).

The experiment took place during the months of October and November 2015. The task consisted of five steps. (1) Firstly, participants connect to a controlled server environment. In this environment, the participants have access to the application that guides them through the experiment, the necessary datasets, an introductory video, Microsoft Office applications and the open-source tool R. Furthermore, Internet access is provided. (2) Consecutively, each participant has the opportunity to watch a small introduction to analytics and the assignment[4]. This

[1] `http://associatie.kuleuven.be/eng/about`

[2] See Section 1 in the Supplementary Material

[3] `https://azure.microsoft.com`

[4] See `http://www.dataminingapps.com/wp-content/uploads/2015/09/Cluster-English.mp4` (unsupervised problem) and `http://www.dataminingapps.com/wp-content/uploads/2015/09/churn-English.mp4` (supervised problem).

Table 1: Description of the participants' pool.

	Novice	**Expert**
Sample size	82	22
Geographical area	63.75% are Belgian students; exchange students come from 20 different countries from Europe, North America, South America, Asia and Africa	9 different countries from Europe, North America, South America and Asia.
Age	$[20; 37]$; median= 21	$[23; 45]$; median= 30
Gender	60% male	95.45% male
Experience with Azure	Yes: 1.25% No, but experience with similar tools: 8.75% No: 90.00%	Yes: 4.55% No, but experience with similar tools: 31.82% No: 63.63%
(Previous) education.	Undergraduate: 33.75% Graduate: 66.25%	Undergraduate: 13.64% Graduate: 68.18% PhD: 18.18%
	Marketing: 23.75% Business & economics: 26.25% IT: 18.75% Finance: 13.75% Statistics: 6.25% Other: 11.25%	Data Science: 36.36% Engineering: 22.73% IT: 13.64% Other: 27.27%
Work experience / Business domain	32.50% has work experience	Analytics: 63.64% Risk management: 22.73% IT: 4.55% Finance: 4.55% General management: 4.55%
Number of statistics tools and programming languages with which you have experience (out of 14)	Mean: 2.78 Median: 1 Range: $[0, 11]$	Mean: 4.78 Median: 5 Range: $[2, 14]$
MCQ Test (out of 5)	Mean: 2.14 Median: 2	Not applicable

presentation is motivated by the assumption that each employee who performs analytics, will have received at least a small introduction to analytics and the employed platform. (3) The next step exists of a small pre-experiment questionnaire[5]. By means of multiple-choice questions, we collect demographics and information about existing knowledge and experience, adapted to the target group (novice or expert). (4) The participants are then guided to Azure in order to perform the analytics task. They are also requested to answer some questions with regards to the performance of their solution. (5) Finally, a small post-experiment questionnaire[6], inquires all participants about their user experience. During the whole process, the screen of each participant was monitored using the tool Procrastitracker[7].

3.2 Experimental design

The experiment is designed as a full 2^3 factorial design with factors *expertise level*, *problem setting* and *data quality*. Figure 1 illustrates the different components of the methodology. We take the expertise into account as both novices and experts are included. Next, the participants either perform a supervised or unsupervised

[5] See Section 2 in the Supplementary Material

[6] See Section 3 in the Supplementary Material

[7] `http://strlen.com/procrastitracker`

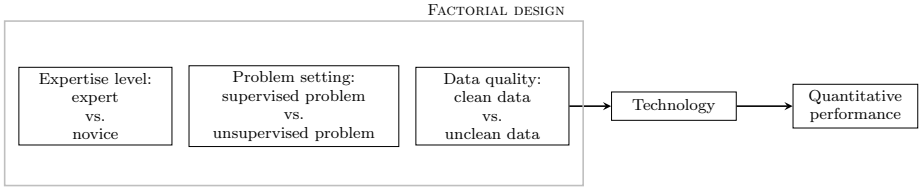


Fig. 1: Research methodology components.

analytics task on either a clean or an unclean dataset. Both the type of task and whether they received a clean or unclean dataset, were assigned randomly and evenly to the participants.

We specifically choose to include both a supervised and unsupervised problem as factor levels, since most analytical problems fall into one of these categories. Performance is measured by means of multiple metrics, see Appendix A for more details. However, we focus on one specific metric for each problem in order to limit redundancy. Firstly, because of its popularity, a churn prediction problem is chosen for the supervised setting. We employ a public dataset available at the UCI library[8]. It consists of 5,000 customers of a telecommunications company with a churn rate of 14.14%; and includes 17 features. We measure performance as the area under the receiver operating characteristic (ROC) curve (AUC), which ranges from 0 to a maximum of 1. This is a popular, well-known metric for binary classification problems in the data analytics community. To this end, participants were required to apply their model on a validation set containing 20% of the observations of the original dataset with omitted labels.

Secondly, we opt for customer segmentation as the unsupervised problem. The dataset[9] for this task was created by the authors. As such, we are able to compare the participants' solutions with the model solution generated in Azure. Four customer profiles are deliberately put in the dataset, while also introducing a number of 'noise' customers in order to increase the credibility of the dataset. The ideal model solution was then generated in Azure, in order to ensure that participants could achieve a perfect solution using the AaaS tool. The dataset consists of 11 features that describe 5,000 customers of a fitness center. Participants were asked to return a clustering solution for the given customer dataset. The retrieved clusters are compared with the actual customer profiles present in the custom-made dataset using the measure *similarity* defined by Gavrilov et al. (2000), see Equation 1. This measures is implemented by Montero & Vilar (2014) and used in various other works (Liao, 2005; Montero & Vilar, 2014).

$$\text{similarity}(G, C) = \frac{1}{k}\sum_{i=1}^{k} \max_{j\in[0,l]} \text{Sim}(G_i, C_j) \text{ with } \text{Sim}(G_i, C_j) = \frac{2|G_i \cap C_j|}{|G_i| + |C_j|} \quad (1)$$

Here, G is the ideal clustering solution, C is the clustering solution of the participant, k is the number of clusters in G, l is the number of clusters in C, and $|\cdot|$ denotes the cardinality of the respective set. Thus, this metric assumes that a 'ground-truth' clustering solution (G) exists (Montero & Vilar, 2014) to which

[8] http://www.sgi.com/tech/mlc/db

[9] http://www.dataminingapps.com/customer-segmentation-fitness/

the participants' solutions (C) are compared. For every cluster in G, the metric selects the most similar cluster from the participant's solution. Consecutively, these similarities are summed. Note that similarity of clusters is calculated by taking the intersection of both clusters and adjusting this number for the total amount of customers in both clusters. Inherent to its definition, this metric more closely resembles classification metrics compared to typical distance-based clustering measures. A second advantage of this metric is the fact that the number of clusters in G and C should not be the same in order to apply the metric. Participants can still generate solutions that translate well to the actual profiles although they have a different number of clusters than the ideal solution. Moreover, observations for which a segmentation label is lacking are all clustered together and treated as a separate cluster. Finally, we normalize both the AUC and the similarity metric to $[0; 1]$ according to the best-performing participant (in terms of the relevant metric), see Equation 2. This normalization procedure was applied in order to improve comparability between metrics. It can be assumed that the participants are not able to deliver a perfect model nor would this be desired. By normalizing their scores, we can compare their performance with regards to the maximum possible performance achieved by novices and experts. Note that for both performance metrics, a higher score indicates a better performance.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

Next, we distinguish between a clean and unclean dataset, because of its impact on the performance of analytics models (Moges et al., 2016), commonly referred to as the 'garbage in, garbage out' principle. We simplify the definition of an unclean dataset to a dataset that contains errors such as missing values and outliers. As such, for both the supervised and unsupervised settings, approximately 1% of observations in the unclean dataset were converted to missing values, and 0.12% of the data are transformed to outliers. These parameter values are determined by aiming for a balance between the impact of data quality on the analytics solution and the required effort of preprocessing the dataset.

3.3 Data analysis

We extract several variables from the experiment, which we then analyze by means of a factor analysis, linear regression and process analysis. The first technique aims to gain insights into the three main experiment factors and their relation to analytics task performance. The latter two techniques zoom in the novices and on the importance of their characteristics, the task characteristics and the approach followed.

3.3.1 Variable extraction

We identify three types of variables. Firstly, we can use the variables of Table 1, all related to the individual user ('UserVar'). Next, we can define variables based on the task ('TaskVar'), i.e. the problem setting and data quality. Finally, we have variables related to the approach ('ApproachVar') the participant took. These are extracted based on three sources. We transfer the information from the

Table 2: Event log example of the behavior of novices.

User ID	Activity	DateTime	Program	Task category	Performance
user1	churn-English.mp4	2015-10-13 14:02	MediaPlayer	WatchPresentation	0.56
user1	Experiments	2015-10-13 14:08	Azure	BuildModel	0.56
user1	split in microsoft azure	2015-10-13 14:49	Google	Documentation	0.56
		...			
user21			...		

post-experiment questionnaire into variables representing the participants' perception of how much they used the available tools and which steps of the analytics process (Fayyad et al., 1996) they followed. Furthermore, we analyze the actual final model of the participants in Azure in terms of modules used and whether these modules belong to visualization, data preprocessing, data transformation, data mining, model evaluation or model interpretation. Finally, we use data on the actual logged behavior. This data is transformed into event logs, as illustrated in Table 2. An event log displays each occurrence of an activity, and adds a participant identifier, a time element and a resource, which in this case is the program used. In addition, we kept track of the duration of each event, both in terms of total number of seconds and total number of active seconds. Furthermore, we categorize each activity in an aggregated task and label the logs accordingly. This leads to a total of 20 task categories. On the whole, 73 relevant variables are extracted. A full overview of the variables can be found in Table C.1 in Appendix C.

3.3.2 Data analysis techniques

Factorial analysis Firstly, we analyze how the three factors —expertise level, problem setting, and data quality— affect the analytics task performance in terms of *AUC* or *similarity* by means of a full factorial analysis. We apply analysis of variance (ANOVA) on (1) the original dataset and (2) the aligned rank transform (ART) dataset. Applying a rank transform method, is recommended in cases where the strict assumptions of ANOVA are not fulfilled (Conover et al., 1981). For this reason, researchers turn to non-parametric analysis. However, conventional RT methods were found to be only accurate for estimating main effects, not the interaction effects between factors (Wobbrock et al., 2011). The ART method, on the other hand, firstly aligns the response variable according to the effect of interest before the response is ranked, thereby addressing this limitation. This is consecutively repeated for each effect of interest, including the interaction effects. For more information, we refer to Wobbrock et al. (2011). We specifically include both a parametric ANOVA and an ANOVA on ART data (ART-ANOVA), because not all assumptions of ANOVA are supported by our dataset. As such a normal distribution of the residuals is rejected by the tests of Anderson & Darling (1954); Jarque & Bera (1987) and Shapiro & Wilk (1965) on a 5% significance level. Moreover, homoscedasticity is rejected on a 5% significance level by the tests of Bartlett (1937) and Fligner & Killeen (1976) but not by the test of Levene (1960).

Linear regression analysis Secondly, we take a closer look at how the characteristics of the user, the task and the approach that the user followed, impact analytics performance. Beforehand, we perform an initial feature selection by means of a correlation analysis, based on Pearson correlation coefficient with a cut-off of 0.5,

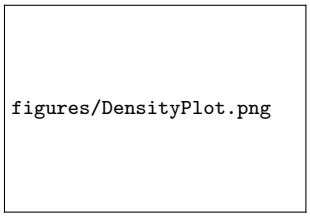


Fig. 2: Density plot of $AUC/similarity$ showing three peaks in performance.

Pearson's chi-squared test with a cut-off p-value of 0.05 and variance inflation factors (VIF) with a cut-off threshold of 4. This procedure leads to a further reduction of the set of variables to 41 external factors. Next, with our reduced set of variables, we build a linear regression model with a stepwise forward and backward feature selection based on the Akaike information criterion (AIC), in order to explain the performance of participants.

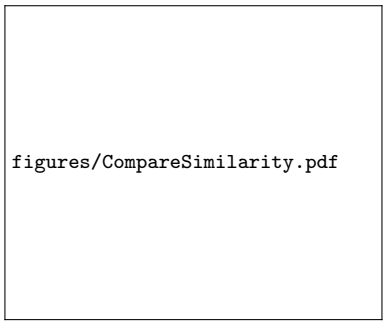
Process analysis Finally, we also visualize the workflow of the participants to gain additional insights into their approach. Event logs, such as in Table 2, can be used to identify trends and patterns by applying process analysis techniques. The tool Disco[10] was used to construct a visual representation of user workflows which allows for further inspection and analysis. This visualization is also called a process map. A process map shows the different traces that occur in the event log. Each trace follows one particular participant throughout the whole experiment. As such, we can visually assess which paths are frequently followed and how participants navigate through the task at hand. For this purpose, we divide the participants in four groups: experts, high-performing novices (with performance $\in]0.7, 1]$), medium-performing novices (with performance $\in]0.4, 0.7]$), and low-performing novices (with performance $\in [0, 0.4]$). This allows us to compare the behavior of novices to that of experts based on three levels of performance. We explicitly split the novice group in three based on the density plot of $AUC/similarity$, see Figure 2, which shows three clear peaks in performance. For more information on process analysis in general, the reader is referred to van der Aalst (2011).

4 Results and discussion

4.1 A general comparison of the achieved performance with an AaaS platform

Firstly, we examine the performance of both novices and experts using AaaS, compared to random baseline models. Figure 3 illustrates two aspects of user performance. Figure 3a compares the customer segmentation models of novices and experts to a random customer segmentation solution. The random model is created by randomly assigning the customers to four equally distributed segments. For the supervised analytics task, the AUCs of novices and experts are compared with a random model with AUC equal to 0.50 (normalizing the AUC by the best-performing participant gives a score of 0.5365), see Figure 3b. Firstly, we

[10] https://fluxicon.com/disco



(a) Unsupervised task

(b) Supervised task

Fig. 3: Comparison between expert, novice and random performance by means of (a) contrasting *similarity* between a random clustering into four clusters, novices and experts; and (b) contrasting the AUC of novices and experts with a random churn prediction model.

observe that while experts still outperform novices significantly for both tasks (p-values < 0.003 and < 0.03 using Student's t-test for the unsupervised and supervised analytics task respectively), we can conclude that novices are empowered by means of AaaS as they are still able to outperform a random solution. The novice group scores significantly better than the random baseline models for both the customer segmentation and churn problem settings (p-value < 0.002 for both tasks, measured by means of Student's t-test). By using this platform, they are able to already greatly improve their performance compared to a random solution, even if the user does not have a background in analytics. AaaS therefore allows to perform analytics task decently, regardless of the expertise level of the user. Although performing better than a random model might seem straightforward, performing an analytical task successfully is already challenging for a novice. Delivering a sufficient result can, as such, have concrete advantages in practice. AaaS might encourage and guide business users in analytics tasks. Nevertheless, we cannot make assumptions on the impact of AaaS compared to other standalone tools or platforms in the cloud. Secondly, we notice a difference in average performance between the two problem settings, both for experts and novices. The supervised task leads to higher normalized scores for both levels of expertise, which indicates that this factor should be examined more closely. Furthermore, we observe that the gap between the average performance of experts and novices is also larger for the unsupervised setting, which suggests that supervised tasks are more approachable.

A factor analysis of the expert level, problem setting and data quality, gives us more information about the impact of each factor on task performance. For this analysis, 82 novices and 22 experts are included (Factor *expertise level*). Out of this population, 53 handled the supervised problem and 51 handled the unsupervised problem (Factor *problem setting*). In total, 52 participants had an unclean dataset and 52 had a clean dataset (Factor *data quality*). This section discusses our findings with regards to the $AUC/Similarity$ performance metric. A generalization to other performance metrics (see Appendix A) can be found in Appendix B. They, in general, confirm the findings from the analysis represented here.

Table 3: Factor analysis for *AUC/similarity*. Significance is calculated using (1) ANOVA and (2) ART-ANOVA. $\hat{b}_i$ represents the estimated effect of factor i and $\hat{b}_0$ equals the estimated mean performance. Factor A defines the *expertise level*, Factor B the *problem setting*, and Factor C the *data quality*.

	Estimate	ANOVA p-value	ART-ANOVA p-value
$\hat{b}_0$	0.6426	NA	NA
$\hat{b}_A$	0.09785	< 0.003	< 0.001
$\hat{b}_B$	-0.16361	< 0.0001	< 0.0001
$\hat{b}_{AB}$	0.01787	0.7724	0.5746
$\hat{b}_C$	-0.002567	0.5714	0.1079
$\hat{b}_{AC}$	0.01247	0.7055	0.3770
$\hat{b}_{BC}$	-0.01049	0.7129	0.6018
$\hat{b}_{ABC}$	-0.0003841	0.9911	0.9964

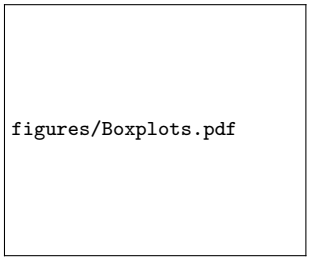


Fig. 4: Boxplots of *AUC/similarity* performance according to the factors *expert level*, *problem setting* and *data quality*.

As can be seen in Figure 4, experts show a higher average performance than novices, while customer segmentation displays worse average performance results compared to churn prediction. In addition, participants with clean datasets obtain on average better performance than users with unclean datasets. These observations are supported by the results of the full factorial analysis, as can be observed in Table 3. We find that only factors *expertise level* and *problem setting* are strongly significant, which means that experts perform significantly better than novices and that churn prediction receives significantly better performance scores than customer segmentation. The relationship between the significant factors is given by Equation 3.

$$\text{Task performance} = 0.6426 + 0.09785x_A - 0.16361x_B + \varepsilon, \tag{3}$$

with

$$x_A \in \{-1; 1\} = \{\text{novice; expert}\}$$

and

$$x_B \in \{-1; 1\} = \{\text{customer segmentation; churn prediction}\}$$

indicating the factor levels for Factors A *expertise level* and B *problem setting*, respectively, and ε as a random error factor.

These results establish a significant difference in user performance depending on both the user and the task characteristics. It is important to note, moreover, that the interaction between Factors A and B is not significant. Therefore, the

problem setting itself has a large impact on the performance of a user, regardless of their level of expertise in data science. Furthermore, we can also observe that the coefficient of Factor B in Equation 3, is twice as large as the coefficient of Factor A, leading to a difference of 33 percentage points between both problem settings and 20 percentage points between both expertise levels. This task characteristic of unsupervised versus supervised problems therefore has a much larger impact on user performance when using AaaS than the level of expertise of the user.

4.2 Extended analysis of novice performance

Apart from the analysis of the main effects described in the previous subsection, additional analyses were carried out to better understand the behavior of novices on their model performance by means of the variables described in Section 3.3.1.

4.2.1 Linear regression

A linear regression is applied to study the correlation between, on the one hand, the user, task and user's task approach characteristics and, on the other hand, the performance of the novices measured by $AUC/similarity$. A few additional novices were excluded from these analyses due to missing values in either their questionnaires or process tracking, leading to a total of 71 novices. All numeric variables were normalized in order to easily perceive the relative influence of the significant variables in their coefficients. The final model has an adjusted R^2 value of 0.58. All variables that proved to be significant at the 95% confidence level, are summarized in Table 4.

Firstly, the results of the linear regression confirm the full factorial analysis, as the problem setting is the only task characteristic that has a significant influence on the performance. The customer segmentation problem has a highly negative impact on the final performance.

In terms of user characteristics, two variables remain after the feature selection process: nationality and work experience. Nationality was expressed in a binary variable, as most students have the Belgian nationality (46 out of 71 students). The results indicate that having a different nationality has a negative effect on performance. This effect might be attributed to a larger variance in previous educational background, as the students originate from 21 different countries. Secondly, previous work experience seems to have a large positive effect on the final result of the novices. This result might be related to inherent qualities of novices who have already worked, such as a higher maturity level and a better general understanding of the business relevance of the problem settings.

Finally, the user task approach seems to be vital for the success of a novice using AaaS with a total of 10 significant variables. With regards to the user perception of the time spent on a certain step in the analytics process, the visualization, transformation and actual model building steps are all positively correlated with the performance, while more time spent on data preparation has a negative impact on the final performance. However, the actual number of modules in the final model that transform the data has a negative coefficient. This indicates that deliberating longer on which transformation to apply, has a positive influence on the final result, while simply applying more transformations does not necessarily lead to a better

Table 4: Additional analysis for novices with *p-value < 0.05; **p-value < 0.01; ***p-value < 0.001. The table excludes variables that are part of the linear regression, but are not significant at the 95% level. These non-significant variables include gender, whether or not the novice expects to do similar exercises in their future job, the number of modules used for data mining and the number of programs used during the experiment.

Type of variable	Variable	Estimate	Std. error	p-value
Task	Cluster	-0.22337	0.07029	0.002474**
User	Non-Belgian nationality	-0.16371	0.07506	0.033631*
	Work Experience	0.19873	0.07056	0.01065*
User Approach	Perceived time for Visualization	0.45869	0.19208	0.020534*
	Perceived time for Data Preparation	-0.38943	0.17691	0.032095*
	Perceived time for Transformation	0.79737	0.17425	< 0.0001***
	Perceived time for Data Mining	0.34158	0.15503	0.031946*
	Number of modules for Preprocessing	0.44916	0.16075	0.007229**
	Number of modules for Transformation	-0.65397	0.15590	0.000104***
	Number of activities	0.30476	0.14206	0.036538*
	Internet Search	-0.41806	0.17690	0.021815*
	View slides	-0.47799	0.21756	0.032409*
	Watch presentation	-0.43169	0.12877	0.001484**

model. Contrastively, the number of modules that focus on the preprocessing of the data, such as removing missing values, has a positive effect on the performance, while the perceived time spent on the same activity has a negative coefficient. This effect is due to the fact that data cleaning ultimately does have a positive impact on the final result, but the amount of time that is spent on reading in and preprocessing the data takes away from other important steps in the analytics process. Finally, the variables collected by means of Procrastitracker, and therefore related to the actual process approach of the novices, prove to be very important. In terms of the number of different activities that show up in their work flow, the novices with a more complex process perform better. When looking into the nature of the activities, novices who spend more time on Internet search and on reviewing the slides and presentation that were provided, tend to achieve a lower accuracy. This set of significant variables indicate that the worst performing novices show signs of confusion, as they spend their time looking at the general problem descriptions in the slides/presentation and searching on the Internet instead of experimenting with the AaaS tool and its documentation.

Together, these variables indicate that the performance of novices is mainly explained by the approach that they follow to achieve their results. Furthermore, the problem setting and having previous work experience also have a strong impact on the final performance.

4.2.2 Process analysis and visualization

For the process analysis, we analyze in total 76 novices and 12 experts. The performance of the novices is categorized into 15 low-, 24 medium-, and 37 high-performing novices. We found that experts worked in total significantly longer, i.e. 138 minutes, than novices, i.e. 101 minutes (p-value < 0.001, MannWhitney U test). These results can be further extended to novices. Namely, high-performing

novices spent more time, i.e. on average 110 minutes, than medium-performing novices, i.e. on average 93 minutes (p-value < 0.01, Student's t-test). Similar conclusions can be drawn for the total number of active minutes, although to a lesser extent. As such, high-performing novices worked actively longer, i.e. on average 122 minutes, compared to medium- and low-performing novices, i.e. both on average 105 minutes (p-values < 0.05, Student's t-test). If we zoom in on novices, lower performance correlates with less unique programs, less unique relevant activities, and less total relevant activities. These insights can also be deduced from the process maps. Figure 5 visualizes the paths of, on the one side, experts and, on the other side, low-performing novices. We can clearly observe that the process of low-performing novices is less structured. Moreover, as we can deduce from the quantitative analysis and the process maps, the higher the performance classification of the novice, i.e. low, medium or high, the more similar their process is to that of an expert. Thus, we can conclude that experts and better-performing novices work longer and have a higher number of activities, take advantage of more programs to solve their task but, nevertheless, follow a more structured and straightforward path. This is in line with the result from the linear regression indicating that better performing novices performed more activities and used more programs.

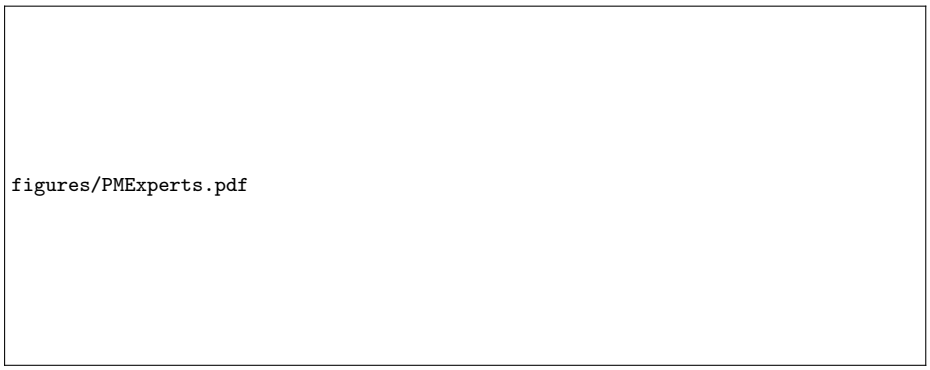
4.3 Discussion

We can conclude from these analyses that AaaS seems to be a useful platform for novices, as users were able to achieve a satisfying performance for both supervised and unsupervised tasks compared to a random baseline model. However, supervised tasks seem to be more approachable for AaaS, which holds for both experienced and inexperienced users. We also found that, although experts outperform novices when using AaaS for analytics tasks, significant differences exist among novices. Firstly, user characteristics, such as work experience, play a role in user performance. This can help managers in their first selection of potential candidates. However, user's task approach characteristics proved to be the most significant explanatory factor in the analyses. Novices that tackle the problem in a rather structured manner, with an approach that is similar to the experts', generally have a more successful outcome. This can contribute to management as well as AaaS vendors. Management can apply these insights when conducting trainings while vendors can design their platforms so the analytics process is optimally supported.

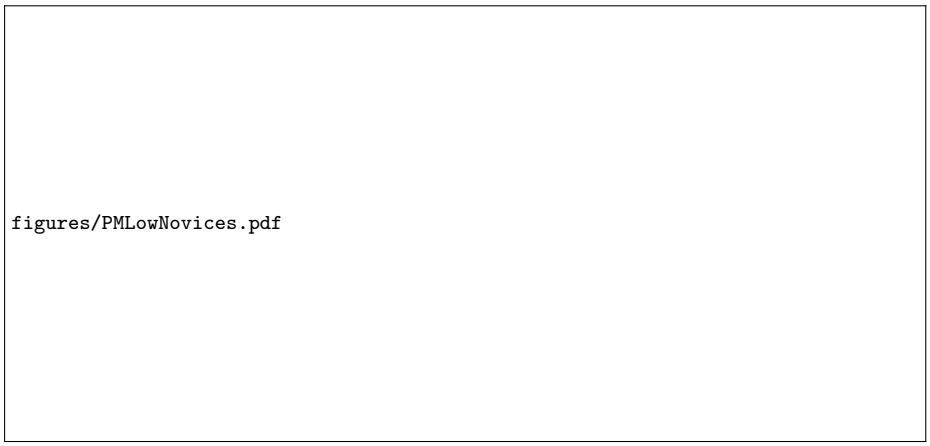
5 Limitations and further research

5.1 Addressed threats to validity

As recommended by Boudreau et al. (2001) and Straub (1989), a *pilot test* was performed to test the instrumentation. Participants provided oral and written feedback during and after the experiment on both content and formulation. Secondly, *content validity* was improved by using both supervised and unsupervised



(a) Experts



(b) Low-performing Novices

Fig. 5: Process maps of the paths (a) experts and (b) low-performing novices follow while solving the analytics task. The activities are categorized. The maps show case frequencies, i.e. how often a participant engaged in an activity at least once. To improve readability, we filtered 10% of the least occurring traces, and focused on activities which at least either one third of the experts or low-performing novices applied.

problem settings, and clean and unclean datasets. The *construct validity* was enhanced by applying multiple performance measures, which were all normalized for comparability. Furthermore, to ascertain that Azure was suitable for the tasks at hand, we compared positive and negative feedback on the service from the participants in terms of their performance. For all novices, a normalized feedback score was calculated by subtracting the number of negative feedback points from the number of positive feedback points and then normalizing this value. Using Pearson's product-moment correlation test, the correlation between the feedback score and the $AUC/similarity$ performance ($\rho = 0.07998$) was not significant (p-value > 0.48). Next, some measures were taken to improve *internal validity*. Participants were, as such, randomly assigned to the factor levels of Factor 'problem setting' and Factor 'data quality' in order to reduce selection bias. More-

over, novices were performing their task in a controlled classroom setting with no inter-participants communication allowed and were given a maximum of three hours to complete the experiment. Both Tukey's Honest Significant Difference test and the Kruskal-Wallis Rank Sum test also indicated that there were no significant differences in performance between groups participating on a different date (p-value > 0.91), on a different day in the week (p-values > 0.98) or at a different time during the day (p-value > 0.96). Finally, the *generalizability* of this study is improved by the size of the sample of novices, which greatly surpasses the mean sample size of 48.6 reported for controlled experiments in software engineering (Sjøberg et al., 2005). Although 'only' 22 experts participated, the use of professionals as experts further increases generalizability.

5.2 Further research

Some threats to validity remain. These limitations are, however, regarded as potential future research rather than as a liability. Firstly, by focusing on one specific platform, a benchmark was deliberately avoided. However, repeating the study with one or multiple other AaaS platforms, would further enhance generalizability. Similarly, the study can be repeated using other supervised and unsupervised problems, such as multiclass classification, credit scoring, forecasting, etc. Finally, this study only focuses on structured data and can therefore be expanded to unstructured data, such as text and video data. All of these extensions to the set-up of the experiment would further improve its generalizability. In terms of evaluation, the validation of clustering solutions is not a straightforward task, as monotonicity, noise, density, sub-clusters, and skewed distributions might impact the clustering validity (Liu et al., 2013). This complicates the comparison between the performance of supervised and unsupervised solutions. Nevertheless, we aimed to address this by focusing on accuracy-based metrics for both the supervised and unsupervised task and by normalizing the metrics. Finally, the sample of participants showed some limitations. A larger sample size of experts would enhance the validity of the findings. In addition, all experts performed their task remotely, which limits the controllability of the experiment. In terms of the novices, the study could also be repeated for novices with different analytics expertise levels. Moreover, experiments could be undertaken to research the approach novices take in solving analytics tasks, given how important this aspect proved to be in this study. Alternatively, a longitudinal study could be performed to study the learning effect of novices.

6 Conclusion

The aim of this study is to determine whether the analytics process could be made more accessible to a larger audience by using a given analytics-as-a-service platform. More precisely, we investigated whether this type of platform is suitable for users with varying levels of expertise and for different analytics tasks. Furthermore, this paper looked into which user, task and user's task approach characteristics influence the performance of novice users. To test this, a full factorial experiment

was designed and a number of inexperienced and experienced people were asked to solve standard analytics tasks using a web-based platform.

The results of the experiments show that, in the context of AaaS platforms, novices are in general able to outperform random benchmark models. Furthermore, there is a significant difference between the two analytics tasks as well, where participants with the churn prediction task performed 33 percentage points better than users with the customer segmentation exercise regardless of their expertise level, as can be observed from Equation 3. Additionally, more extensive analyses on the group of novices confirm the difference between the supervised and unsupervised analytics problem and show that students with work experience perform better. However, the largest group of significant variables refers to the user's task approach, as novices with more elaborate processes who use more resources and spend more time exploring the data, perform a lot better than others. The process of this high-performing group of novices also shows more similarities with the process that the experts undertake than the group of low-performing novices.

It is illustrated that our study is well founded by confronting the limitations of the experiment. As such internal, external, content and construct validity are addressed, as well as the reliability of the results. Out of this overview, possibilities for further research arise. For example, variations with other analytics problems and data can be implemented. Furthermore, the effect of learning could be studied based on the findings of this study.

While data science experts sill achieve the best performance with AaaS platforms, this service does offer a viable analytics solution for business users. Given a suitable task, novices that make use of all of the resources available and know how to structure their approach, deliver better results on average. However, further research would need to be undertaken to assess the impact of learning in this context. Nevertheless, these findings are already interesting for AaaS vendors who wish to further improve their tools. The question therefore remains how extensive training for novices has to be and whether at the end of it, they have not become experts themselves.

Acknowledgments

This work was supported by Colruyt Group; and the Coca-Cola Company.

References

Alpar, Paul, & Michael Schulz 2016. Self-Service Business Intelligence. Business & Information Systems Engineering, 58(2):151–155.

Anderson, Theodore W, & Donald A Darling 1954. A test of goodness of fit. Journal of the American Statistical Association, 49(268):765–769.

Armbrust, Michael, Armando Fox, Rean Griffith, Anthony D Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, et al. 2010. A view of cloud computing. Communications of the ACM, 53(4):50–58.

August, Terrence, Marius Florin Niculescu, & Hyoduk Shin 2014. Cloud implications on software network structure and security risks. Information Systems Research, 25(3):489–510.

Baesens, Bart 2014. Analytics in a big data world: The essential guide to data science and its applications. John Wiley & Sons, Hoboken, NJ.

Bartlett, Maurice S 1937. Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, pages 268–282. http://www.jstor.org/stable/96803.

Boudreau, Marie-Claude, David Gefen, & Detmar W Straub 2001. Validation in information systems research: A state-of-the-art assessment. MIS Quarterly, 25(1):1–16.

Chen, Hsinchun, Roger HL Chiang, & Veda C Storey 2012. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36(4):1165–1188. http://www.misq.org/skin/frontend/default/misq/pdf/V36I4/SI_ChenIntroduction.pdf.

Chen, Pei-Yu, & Shin-Yi Wu 2013. The impact and implications of on-demand services on market structure. Information Systems Research, 24(3):750–767.

Chen, Ying, Jeffrey Kreulen, Murray Campbell, & Carl Abrams 2011. Analytics ecosystem transformation: A force for business model innovation. In SRII Global Conference (SRII), 2011 Annual, pages 11–20. IEEE.

Conover, William J, Mark E Johnson, & Myrle M Johnson 1981. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics, 23(4):351–361.

Davenport, Thomas H. 2014. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Harvard Business Review Press, Boston, MA.

Davenport, Thomas H, & Jeanne G Harris 2007. Competing on analytics: The new science of winning. Harvard Business Press, Boston, MA.

Debortoli, Stefan, Oliver Müller, & Jan vom Brocke 2014. Comparing Business Intelligence and Big Data Skills. Business & Information Systems Engineering, 6(5):289–300.

Demirkan, Haluk, & Dursun Delen 2013. Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. Decision Support Systems, 55(1):412–421.

Elazhary, Hanan 2014. Cloud Computing for Big Data. Technical Report 4, MAGNT Research Report.

Fayyad, Usama, Gregory Piatetsky-Shapiro, & Padhraic Smyth 1996. From data mining to knowledge discovery in databases. AI Magazine, 17(3):37. http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230.

Fligner, Michael A, & Timothy J Killeen 1976. Distribution-free two-sample tests for scale. Journal of the American Statistical Association, 71(353):210–213.

Gartner 2015. Magic Quadrant for Business Intelligence and Analytics Platforms. http://www.gartner.com/technology/reprints.do?id=1-2AD8O9T$&$ct=150223$&$st=sb.

Gavrilov, Martin, Dragomir Anguelov, Piotr Indyk, & Rajeev Motwani 2000. Mining the stock market: Which measure is best? In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, MA, USA, August 20-23, 2000, pages 487–496.

Gower, J. C. 1971. A General Coefficient of Similarity and Some of Its Properties. Biometrics, 27(4):857–871.

Gupta, Prashant, A. Seetharaman, & John Rudolph Raj 2013. The usage and adoption of cloud computing by small and medium businesses. Internationam Journal of Information Management, 33(5):861 – 874.

Imhoff, Claudia, & Colin White 2011. Self-service business intelligence-Empowering users to generate insights. TDWI Best Practice Report. https://www.sas.com/resources/asset/TDWI_BestPractices.pdf.

Jaatun, Martin Gilje, Siani Pearson, Frdric Gittler, Ronald Leenes, & Maartje Niezen 2016. Enhancing accountability in the cloud. International Journal of Information Management, forthcoming.

Jarque, Carlos M, & Anil K Bera 1987. A test for normality of observations and regression residuals. International Statistical Review/Revue Internationale de Statistique, pages 163–172.

Leavitt, Neal 2013. Bringing big analytics to the masses. IEEE Computer, 46(1):20–23.

Levene, Howard 1960. Robust tests for equality of variances1. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, 2:278–292.

Liao, T. Warren 2005. Clustering of time series data—a survey. Pattern Recognition, 38(11):1857–1874.

Lismont, Jasmien, Tine Van Calster, María Oskarsdottir, Jan Vanthienen, Bart Baesens, & Wilfried Lemahieu 2015. API for prediction and machine learning: poll results and analysis. KDnuggets News, 29. http://www.kdnuggets.com/2015/09/api-prediction-machine-

learning-poll-results.html.

Lismont, Jasmien, Jan Vanthienen, Bart Baesens, & Wilfried Lemahieu 2017. Defining analytics maturity indicators: A survey approach. International Journal of Information Management, 37(3):114–124.

Liu, Yanchi, Zhongmou Li, Hui Xiong, Xuedong Gao, Junjie Wu, & Sen Wu 2013. Understanding and Enhancement of Internal Clustering Validation Measures. IEEE Transsactions on Cybernetics, 43(3):982–994.

Marston, Sean, Zhi Li, Subhajyoti Bandyopadhyay, Juheng Zhang, & Anand Ghalsasi 2011. Cloud computing - The business perspective. Decision Support Systems, 51(1):176–189.

Mell, Peter, & Tim Grance 2011. The NIST definition of cloud computing. Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology Gaithersburg. http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

Moges, Helen-Tadesse, Vronique Van Vlasselaer, Wilfried Lemahieu, & Bart Baesens 2016. Determining the use of data quality metadata (DQM) for decision making purposes and its impact on decision outcomes: An exploratory study. Decision Support Systems, 83:32 – 46.

Montero, Pablo, & José A Vilar 2014. TSclust: An R Package for Time Series Clustering. Journal of Statistical Software, 62(1).

Ransbotham, Sam, David Kiron, & Pamela Kirk Prentice 2016. Beyond the hype: the hard work behind analytics success. MIT Sloan Management Review, 57(3). http://sloanreview.mit.edu/analytics2016.

Shapiro, Samuel Sanford, & Martin B Wilk 1965. An analysis of variance test for normality (complete samples). Biometrika, 52(3/4):591–611.

Sjøberg, Dag IK, Jo E Hannay, Ove Hansen, Vigdis By Kampenes, Amela Karahasanovic, Nils-Kristian Liborg, & Anette C Rekdal 2005. A survey of controlled experiments in software engineering. Software Engineering, IEEE Transactions on, 31(9):733–753.

Straub, Detmar W. 1989. Validating Instruments in MIS Research. MIS Quarterly, 13(2):147–169. http://misq.org/validating-instruments-in-mis-research.html.

Van Calster, Tine, Jasmien Lismont, María Óskarsdóttir, Seppe vanden Broucke, Jan Vanthienen, Wilfried Lemahieu, & Bart Baesens 2016. Automated Analytics: The Organizational Impact of Analytics-as-a-Service. In 1st Workshop on Enterprise Intelligence in conjunction with KDD 2016, August 14, San Francisco, CA. Forthcoming, available at https://www.researchgate.net/publication/311576573_Automated_Analytics_The_Organizational_Impact_of_Analytics-as-a-Service.

van der Aalst, Wil 2011. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag Berlin Heidelberg, Schleiden, Germany.

Weinhardt, Christof, Arun Anandasivam, Benjamin Blau, Nikolay Borissov, Thomas Meinl, Wibke Michalk, & Jochen Stößer 2009. Cloud Computing – A Classification, Business Models, and Research Directions. Business & Information Systems Engineering, 1(5):391–399.

Wobbrock, Jacob O, Leah Findlater, Darren Gergle, & James J Higgins 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 143–146. ACM.

Zorrilla, Marta, & Diego García-Saiz 2013. A service oriented architecture to provide data mining services for non-expert data miners. Decision Support Systems, 55(1):399–411.

A Performance Measures

Next to *AUC* and *similarity*, we repeat our full-factorial analysis for several other performance metrics in order to improve generalizability. Depending on the problem setting, different performance measures need to be employed.

Supervised problem We include four performance metrics for the supervised problem, i.e. churn prediction. Firstly, the area under the ROC curve (*AUC*) of each model is measured. Receiver operating characteristic (ROC) curves display the false positive[11] rate versus the true positive rate, so it clearly visualizes the trade-off between these two measures. To quantify the results that are displayed by the curve, the AUCs can be compared. Next, *accuracy* is calculated as the sum of the number of true positives and true negatives divided by the number of customers in the dataset. Thirdly, we calculate the *top-decile lift* which quantifies how well the model recognizes churners. For comparability reasons, top-decile lift is represented as a percentage of the maximum top-decile lift. The maximum top-decile lift occurs when all customers in the top 10% most likely to churn according to the model, are actually churners[12]. Finally, we compute the *average* of these three measures. By definition, all metrics indicate a higher performance for higher scores.

Unsupervised problem We define three performance metrics for the unsupervised problem, i.e. customer segmentation. Firstly, the retrieved clusters are compared with the actual customer profiles which are present in the dataset that was specifically created for this experiment. This comparison also takes into account the number of clusters, as users can generate solutions that still translate well to the actual profiles although they have a different number of clusters than the ideal solution. Moreover, observations for which a segmentation label is lacking are clustered together and treated as a separate cluster. This metric is indicated by *similarity*, based on Montero & Vilar (2014), see Equation 1 in Section 3.2. Secondly, the performance metric *fit* designates the ratio of inter- and intra-cluster distance. Intra-cluster distance averages the mean distance within a cluster while inter-cluster distance averages the mean distance between clusters, by means of the general dissimilarity coefficient of Gower (1971). Thirdly, an *average* performance is also determined for this problem as the average of *similarity* and a normalized *fit*. By definition, all metrics indicate a higher performance for higher scores.

Finally, we normalize all metrics to $[0; 1]$ according to the best-performing participant (in terms of the relevant metric).

B Results of the full factorial analysis performed on all performance measures

This section generalizes the results from the full factorial analysis discussed in Section 4.1 to all performance measures presented in Appendix A.

[11] In this case, positives are regarded as churners, while negatives are regarded as non-churners.

[12] Note that this is applicable if the churn rate reaches 10% or higher. The employed dataset has a churn rate of 14.14%.

When applying the full factorial analysis for the other performance metrics, the results in Table B.1 are obtained. Note that we apply both an ANOVA on (1) the original dataset and (2) the aligned rank transform (ART) dataset. In general, these tests support Equation 3 in Section 4.1, as Factors A and B are almost always indicated as significant. Nevertheless, some tests indicate a significant interaction effect of Factors A and B, A and C, or B and C. Furthermore, the *accuracy/fit* measure is the only measure suggesting the significance of Factor C, where clean datasets have a slightly worse ($\hat{b}_C = -0.002085$) performance than unclean datasets. Moreover, this effect is only significant when performing ART-ANOVA. This result may be due to a lack of suitability of the statistical test for this specific measure. The aligned responses' p-values should be close to 1 (Wobbrock et al., 2011) in order for the ART-ANOVA statistical test to be valid. Factor C aligned by $A : C$ has only a p-value of 0.3064 for this particular measure and may therefore produce inconsistent results in this particular case. Moreover, *fit* is the only performance measure which is not accuracy-based. This might also contribute to deviating results.

Table B.1: P-values of factors across performance metrics using (1) ANOVA and (2) ART-ANOVA with *p-value $<$ 0.10; **p-value $<$ 0.05; ***p-value $<$ 0.01; ****p-value $<$ 0.001.

Performance metrics	ANOVA significant factors	ART-ANOVA significant factors
AUC/similarity	A***, B****	A****, B****
AUC/fit	A**, B***	A*, B****
accuracy/similarity	A**, B****	A*, B****, AB*
accuracy/fit	A*, B**, AC*	B***, C**
top-decile lift/similarity	A**, B****	A**, B****, AB*
top-decile lift/fit	A*, BC*	None
average/average	A**, B**	A**, B***

C Overview of the variables

Table C.1 provides an overview of the 73 extracted variables described in Section 3.3.1.

Table C.1: An overview of the 73 extracted variables as described in Section 3.3.1.

Type	Name	Description
TaskVar	Problem	The problem setting: supervised or unsupervised problem
	DQ	Data quality: clean or unclean data
UserVar	BA.MA	Current education level of participant: Undergraduate or graduate
	Birthyear	The birthyear of the participant
	WorkExperience	Do you have work experience?
	Dual.nationality	Participant's nationality: Belgian or non-Belgian
	ExperienceTool	Does the participant have experience with Azure or similar tools
	Gender	The gender of the participant
	Programming.languages	Number or programming languages which the participant has proficient experience with
	School	Where does the participant study: Belgian university or Belgian college
	Statistical.programs	Number of statistical programs which the participant has proficient experience with
	StatTest	Particpant's score on the MCQ test
	Study.direction	Domain of education program, see table 1
ApproachVar: Questionnaire	AddedValue	Does the participant believe that applying Azure to this business problem adds value for the company?
	SimEx	Would the participant like to perform similar exercises in their job?
	PercentTransformation	Self reported percentage of time spent on transformation
	PercentPreparation	Self reported percentage of time spent on pre-processing
	PercentDM	Self reported percentage of time spent on data mining
	PercentVisualization	Self reported percentage of time spent on data visualization
	PercentEvaluation	Self reported percentage of time spent on evaluation
	PercentInterpretation	Self reported percentage of time spent on interpretation
	Selection.transformation	Self reported number of modules used for transformation
	Preprocessing	Self reported number of modules used for preprocessing
	Mining	Self reported number of modules used for data mining
	Evaluation	Self reported number of modules used for model evaluation
ApproachVar: Azure Modules	Module.Selection.Transformation	Number of Azure modules used for data selection and transformation
	Module.DM	Number of Azure modules used for data mining
	Module.Evaluation	Number of Azure modules used for model evaluation
	Module.Preprocessing	Number of Azure modules used for data preprocessing
ApproachVar: Logged Behavior	TotalUsefulActiveMinutesPerUser	Total actual active duration of relevant activities (in minutes)
	nTaskAct	Total number of unique relevant activities (aggregated by task) performed
	nAct	Total number of unique relevant activities
	nPrograms	Number of programs used during the experiment
	TaskBuildModelSec	Seconds spent on building a model
	TaskBuildModelPerc	Percentage of time spent on building a model
	TaskCalculatorSec	Seconds spent on using the computer's calculator
	TaskCalculatorPerc	Percentage of time spent on using the computer's calculator
	TaskDocumentationSec	Seconds spent in Azure's documentation
	TaskDocumentationPerc	Percentage of time spent in Azure's documentation
	TaskEntertainmentSec	Seconds spent on entertainment such as Facebook, newssites, YouTube
	TaskEntertainmentPerc	Percentage of time spent on entertainment such as Facebook, newssites, YouTube
	TaskErrorSec	Seconds spent on receiving error messages on the internet
	TaskErrorPerc	Percentage of time spent on receiving error messages on the internet
	TaskExperimentSec	Seconds spent in experiment application
	TaskExperimentPerc	Percentage of time spent in experiment application
	TaskNavigationSec	Seconds spent on navigating in browser or programs
	TaskNavigationPerc	Percentage of time spent on navigating in browser or program
	TaskPreExpSurveySec	Seconds spent on the pre-experimental survey
	TaskPreExpSurveyPerc	Percentage of time spent on pre-experimental survey
	TaskProcessDataSec	Seconds spent on preprocessing the data
	TaskProcessDataPerc	Percentage of time spent on preprocessing the data
	TaskRunningSec	Seconds spent on running models in Azure
	TaskRunningPerc	Percentage of time spent on running models in Azure
	TaskSearchSec	Seconds spent on searching on the internet
	TaskSearchPerc	Percentage of time spent on searching on the internet
	TaskSettingsSec	Seconds spent on selecting settings
	TaskSettingsPerc	Percentage of time spent on selecting settings
	TaskStartSec	Seconds spent on starting the experiment
	TaskStartPerc	Percentage of time spent on starting the experiment
	TaskTemplateSec	Seconds spent on working with an Azure template
	TaskTemplatePerc	Percentage of time spent on working with an Azure template
	TaskTutorialSec	Seconds spent on Azure tutorials
	TaskTutorialPerc	Percentage of time spent on Azure tutorials
	TaskUnknownSec	Seconds spent on unknown tasks
	TaskUnknownPerc	Percentage of time spent on unknown tasks
	TaskViewDataSec	Seconds spent on viewing the data
	TaskViewDataPerc	Percentage of time spent on viewing the data
	TaskViewSlidesSec	Seconds spent on viewing the slides
	TaskViewSlidesPerc	Percentage of time spent on viewing the slides
	TaskWatchPresentationSec	Seconds spent on watching the presentations
	TaskWatchPresentationPerc	Percentage of time spent on watching the presentations
	TaskNotRelevantSec	Seconds spent on non relevant tasks
	TaskNotRelevantPerc	Percentage of time spent on non relevant tasks

Figures

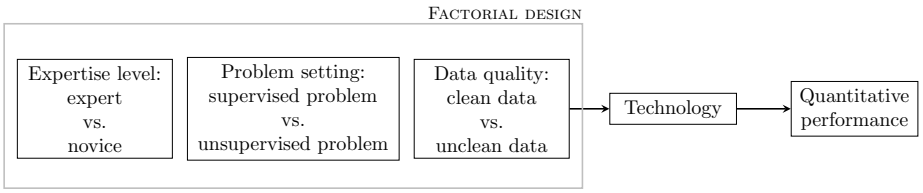


Figure 1: Research methodology components.

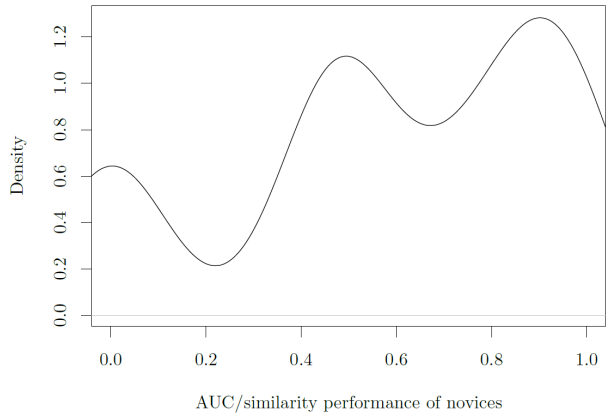
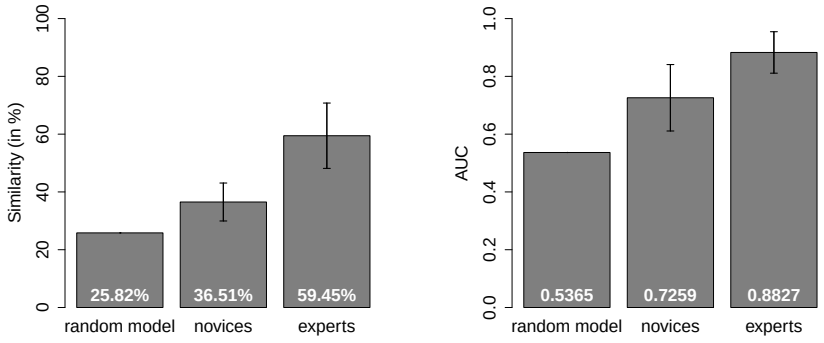


Figure 2: Density plot of $AUC/similarity$ showing three peaks in performance.



(a) Unsupervised task

(b) Supervised task

Figure 3: Comparison between expert, novice and random performance by means of (a) contrasting *similarity* between a random clustering into four clusters, novices and experts; and (b) contrasting the AUC of novices and experts with a random churn prediction model.

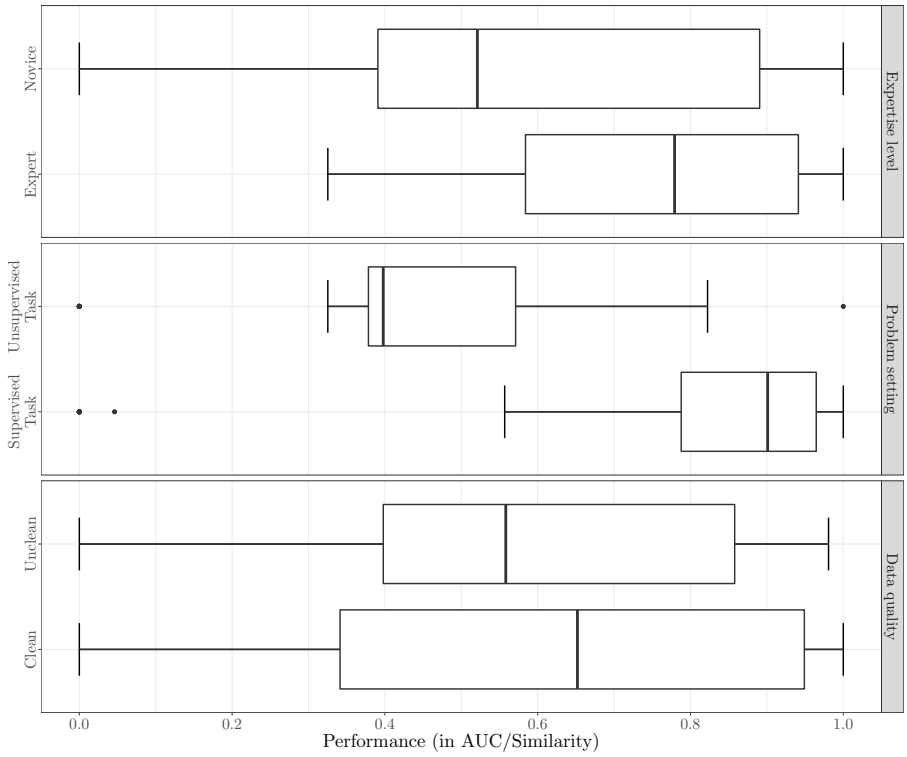
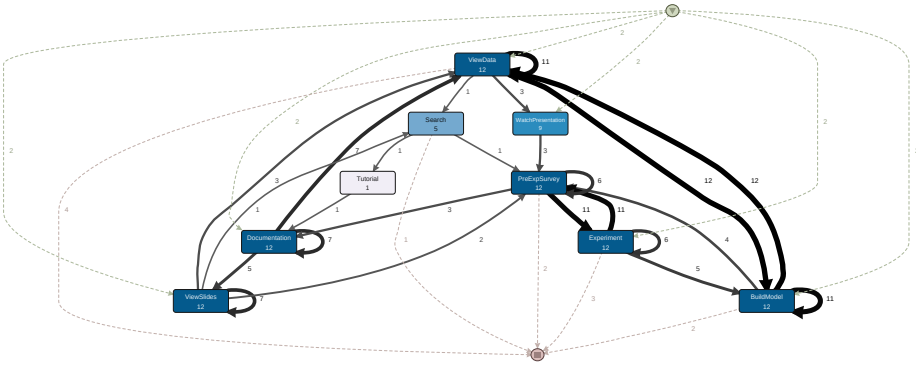
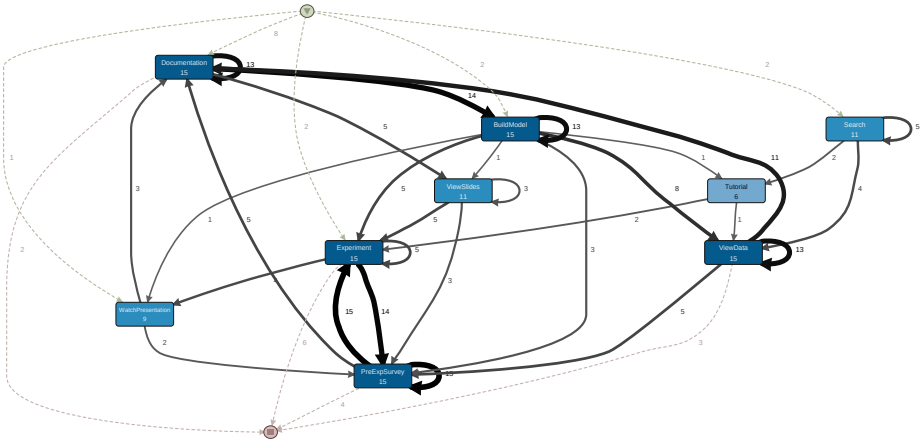


Figure 4: Boxplots of *AUC/similarity* performance according to the factors *expert level*, *problem setting* and *data quality*.



(a) Experts



(b) Low-performing Novices

Figure 5: Process maps of the paths (a) experts and (b) low-performing novices follow while solving the analytics task. The activities are categorized. The maps show case frequencies, i.e. how often a participant engaged in an activity at least once. To improve readability, we filtered 10% of the least occurring traces, and focused on activities which at least either one third of the experts or low-performing novices applied.