https://doi.org/10.48009/4_iis_2019_130-136

Issues in Information Systems
Volume 20, Issue 4, pp. 130-136, 2019

LEARNING OUTCOME EFFECTIVENESS: OPEN-SOURCE VERSUS COMMERCIALLY-LICENSED DATA MINING SOFTWARE TOOLS

Gary Alan Davis, Robert Morris University, davis@rmu.edu
John C. Stewart, Robert Morris University, stewartj@rmu.edu
Diane A. Igoche, Robert Morris University, igoche@rmu.edu

ABSTRACT

Data Analytics, Data Mining, and Machine-Learning have become increasingly prevalent in higher-education, in both on-ground and online curricula. However, few studies have thoroughly and objectively studied the effectiveness of specific Data Mining software (both commercially-licensed and open-source), on learning outcomes. The current study used the Independent Samples T-Test to objectively compare two popular data mining software tools. Specifically, undergraduate, and graduate Information Systems students enrolled in an Introduction to Data Mining course were tested on the use of an open-source data mining tool (i.e., R Programming Language), and on a commercially-licensed data mining tool (i.e., XLMiner from Frontline Solvers). The current study assessed the impact of both tools on undergraduate and graduate student learning outcomes. The findings of this study will be of interest to higher-education faculty who are using, or who may consider using, data mining and machine-learning software in their undergraduate or graduate Information Systems Curriculum.

Keywords: Data Analytics, Data Science, Data Mining, Machine-Learning, Higher Education, Learning Outcomes, Data Analytics Curriculum, Information Systems Curriculum

INTRODUCTION

Although not in widespread use, terms such as *Data Science* have been around since the 1960s. Since that time, some professionals in the computing fields have used "data science" as an alternative to "computer science" (Naur, 1974). More recently, however, Data Science has been defined as "a set of fundamental principles that support and guide the principled extraction of information and knowledge from data" (Provost and Fawcett, 2013 p. 52). A similar term that has evolved in the last decade is *Data Analytics*. Authors Aasheim, et al. define Data Analytics as extending the focus of traditional statistical analysis to the " . . . analysis of larger data sets gathered from a wide variety of data sources" (Aasheim, Williams, Rutner, & Gardiner, 2015, p. 104). Finally, *Data Mining* and *Machine-Learning* have become important components of both Data Science and Data Analytics. Specifically, Data Mining " . . . is the process of discovering patterns in data sets with artificial intelligence, machine-learning, statistics, and database systems" (Barlas, Lanning, & Heavey, 2015, p. 234). The popularity and widespread use of Data Mining and Machine-Learning techniques within business and industry have been enabled by lower data storage costs and more powerful computer processing,

As in business, academia has taken advantage of the lower storage costs and increased processing power to incorporate data mining and machine-learning into their curricula. Similar to business, higher education must make a fundamental choice between commercially-available data mining tools, and open-source software tools. Unlike business, however, institutions of higher education must also be cognizant of the impact of such tools on student learning outcomes. Therefore, it is necessary and appropriate that such tools be reviewed in terms of their effectiveness and impact on learning outcomes.

The purpose of the current study was to compare an open-source data mining package to a commercially-licensed package in order to evaluate each package's effectiveness in increasing learning outcomes. For this research, a commercially-licensed data mining package (i.e., XLMiner from Frontline Solvers) was compared with an open-source data mining package (i.e., R Programming Language). The results of the study will be of interest to faculty who wish to incorporate machine-learning concepts and data mining tools into their Information Systems curriculum.

Specifically, the study sought to answer the following research questions:

RQ1. Is there a statistically-significant difference between an open-source data mining tool (i.e., R) and a commercially-licensed data mining tool (i.e., XLMiner), in terms of learning outcomes effectiveness?
RQ2. Is there a statistically-significant difference between an open-source data mining tool (i.e., R) and a commercially-licensed data mining tool (i.e., XLMiner), in terms of *theoretical concept learning?*
RQ3. Is there a statistically-significant difference between an open-source data mining tool (i.e., R) and a commercially-licensed data mining tool (i.e., XLMiner), in terms of *practical skills learning?*
RQ4. Out of the three learning types (i.e., *Theoretical Concept, Practical Skills,* or *Overall*), which learning type is impacted the most by the choice of data mining tool?

REVIEW OF LITERATURE

Various studies have examined both commercially-available and open-source data mining software tools. Some researchers have focused specifically on open-source tools in their studies. For example, Pushpalatha, et al. reviewed Clementine, RapidMiner, Statistical Analysis System (SAS), and R. Although the authors explored each tool in terms of benefits and features, no recommendations were made regarding which tool is superior for classroom use (Pushpalatha, Saravanan, & Ranjithkumar, 2014). In addition, the authors did not evaluate the tools in terms of each tool's possible impact on learning outcomes. Barlas, Lanning, and Heavey also reviewed open-source data mining tools. In their comprehensive study, the authors evaluated 70 open-source data mining tools across 64 criteria. In their study, Barlas, et al. did recommend the following five open-source tools: RapidMiner, R, Weka, and two Python libraries (i.e., RPy2 and Scikit-learn). While specific tools were recommended based on " . . . the widest variety of data mining tasks and methods," the authors, like the previous study, did not examine the tools in terms of learning effectiveness or outcomes (Barlas, Lanning, & Heavey, 2015, p. 246).

Some authors included both open-source and commercially-available data mining tools. In their study, Haughton, et al. reviewed three commercially-licensed data mining packages (i.e., SAS Enterprise Miner, SPSS Clementine, and XLMiner), and two open-source packages (i.e., GhostMiner and Quadstone). Regarding the commercially-licensed data mining tools, the authors found SAS Enterprise Miner to be superior to the others evaluated, in terms of completeness and capabilities. The authors seemed to be torn between the two open-source offerings, however, they seemed to favor Quadstone over GhostMiner because of Quadstone's ability to handle very large datasets (Haughton, et al., 2003). As in the previously cited studies, no evaluation was made regarding the tools' potential impact on the learning outcomes of students.

Wang, et al. also examined both commercially-licensed and open-source data mining tools. However, in their 2008 study, the authors evaluated each data mining tool according to its appropriateness in relation to five "business/organizational scenarios." One "scenario" in the study, for example, involved a small start-up landscaping company with less than 50 employees. "Educational Institution" was also included as one of the five scenarios in the Wang, et al. study. Specifically, in the study, Microsoft SQL and GhostMiner ranked first and second, respectively, for the "Educational Institution" scenario (Wang, Hu, Hollister, & Zhu, 2008). Again, no evaluation was made as part of the study regarding the impact on learning outcomes or learning effectiveness.

Finally, Zhang and Segall (2010) not only looked at both commercially-licensed and open-source data mining tools, but also extended their study to include text and web mining software tools as well. In terms of the highest-ranked, commercially-licensed data mining tool, the authors recommended SAS Enterprise Miner (consistent with other evaluations of commercially-licensed data mining software). In terms of open-source tools, the authors ranked Megaputer PolyAnalyst as the highest. Even though their study was more comprehensive than others cited, the Zhang and Segall study did not evaluate the data mining tools in respect to student learning outcomes.

As outlined in the previous sampling of studies, data mining software tools have been extensively evaluated in academic literature. As previously indicated, the studies have encompassed both commercially-licensed data mining tools, and open-source tools. Some studies have even examined text and web mining software tools, in addition to data mining tools. However, the existing body of academic research regarding the impact of these tools on learning outcomes is insufficient. The current study sought to address this shortcoming in the existing literature.

RESEARCH METHODOLOGY

This study was conducted to determine if a significant difference in knowledge transfer and the ability to apply machine-learning concepts was affected by the type of data mining software tool used. More specifically, the study asked if there is a difference in the use of a commercially-licensed software package (i.e., XLMiner), versus a code-based, open-source tool with a steeper learning curve (i.e., R Programming Language), in terms of learning outcomes? The research participants in the study were undergraduate and graduate students who were enrolled in different sections of an *Introduction to Data Mining* course.

This study involved the in-class testing of each student participant in each course section. A total of 75 student participants were divided into two sections of the course that used XLMiner as the data mining software tool, and two sections of the course that used R as the data mining software tool. A total of 43 graduate and upper-level undergraduate students used XLMiner, while 32 upper-level undergraduate and graduate students utilized R as their software tool for the course. At the end of the course, each student took a 20-question, multiple-choice test comprised of the same ten machine-learning *theoretical concept* questions, and ten *practical skills* questions that were specific to the software tool used in their section of the course. A standard, timed, online, multiple-choice test format was used for assessing the student participants. The results of both *practical applied learning* of the software, and *theoretical concept learning* were assessed separately, and then combined to determine an overall measurement of the level of knowledge transfer and learning via the use of the two software tools.

The Independent Samples t-test was used to assess the difference in understanding machine-learning concepts, comprehension of the software tool, and overall test scores between the XLMiner and R sections of the course. As mentioned previously, the results of both *practical software skills* and *theoretical concept learning* were assessed separately (using the Independent Samples t-test), and then combined to determine an overall measurement of the level of knowledge transfer and learning.

RESEARCH PARTICIPANTS

As stated previously, a total of 75 students participated in this study. The research participants were comprised of undergraduate and graduate students enrolled in multiple sections of an *Introduction to Data Mining* course. The students in each class section completed an online test, which was comprised of 20 multiple-choice questions. Within the group of student participants, slightly over two thirds were male (69.3%) and slightly less than a third (30.7%) were international students. Finally, 42.7% of the participants were undergraduate students; the remaining 57.3% were graduate students. **Table 1: Participant Demographics by Sex**, and **Table 2: Participant Demographics by Country of Origin** show the demographic breakdown of the research participants.

Table 1. Participant Demographics by Sex
Demographic Breakdown of Research Participants by Class Section and Sex

Class Sections	**Sex**		**Total Students (Cumulative %)**
	Male	**Female**	
Undergraduate Sections - R Programming	7 (70.0%)	3 (30.0%)	10 (13.3%)
Graduate Sections - R Programming	14 (63.6%)	8 (36.4%)	22 (42.7%)
Undergraduate Sections - XLMiner	17 (89.5%)	2 (10.5%)	19 (68.0%)
Graduate Sections - XLMiner	14 (58.3%)	10 (41.7%)	24 (100.0%)
Total	**52 (69.3%)**	**23 (30.7%)**	**75 (100.0%)**

Table 2. Participant Demographics by Country of Origin
Demographic Breakdown of Research Participants by Class Section and Country of Origin

Class Sections	Country of Origin Domestic	International	Total Students (Cumulative %)
Undergraduate Sections - R Programming	9 (90.0%)	1 (10.0%)	10 (13.3%)
Graduate Sections - R Programming	20 (90.9%)	2 (9.1%)	22 (42.7%)
Undergraduate Sections - XLMiner	11 (57.9%)	8 (42.1%)	19 (68.0%)
Graduate Sections - XLMiner	12 (50.0%)	12 (50.0%)	24 (100.0%)
Total	**52 (69.3%)**	**23 (30.7%)**	**75 (100.0%)**

RESULTS

In order to address the first research question, "Is there a statistically-significant difference between an open-source data mining tool (i.e., R) and a commercially-licensed data mining tool (i.e., XLMiner), in terms of learning outcomes effectiveness?," the participants answered a 20-question, multiple-choice test that assessed both *theoretical concept* learning and *practical skills* learning. The assessment data resulting from both the theoretical concept questions and practical skills questions were combined and an Independent Samples t-test (α=0.05) was performed on the entire data set. **Table 3: T-Test Results - Overall Learning** shows the t-test results of the combined data.

As displayed in **Table 3**, there was a *slight* difference in participant learning between the R Programming class sections (M=65.86, SD=14.18) and the XLMiner class sections (M=71.51, SD=20.78) results. However, as indicated by a t-statistic of -1.40, the difference in overall learning between the two software tools was not at a level that was statistically-significant, t(73)=-1.40, p=.166.

Table 3. T-Test Results - Overall Learning
Independent Samples T-Test Results of Combined Practical Skills and Theoretical Concept Learning by Participants

Overall Learning (both Practical and Theoretical Learning combined)	Mean	N	Std. Deviation	t	df	Sig.
R (open-source)	65.86	32	14.18	-1.40	72.53	.166
XLMiner (commercially-licensed)	71.51	43	20.78			

The second research question proposed was, "Is there a statistically-significant difference between an open-source data mining tool (i.e., R) and a commercially-licensed data mining tool (i.e., XLMiner), in terms of *theoretical concept learning?"* In order to address this research question, only the results from the *theoretical concept* test questions were analyzed.

Table 4: T-Test Results - Theoretical Concept Learning shows the comparison of average scores for both R Programming and XLMiner class sections, from the theoretical concept comprehension portion of the assessment. Again, the mean scores between R Programming (M=67.26, SD=16.64) and XLMiner sections (M=76.05, SD=23.01)

show a *slight* difference in participant learning. Although the difference between the software tool results *approached* statistical significance, the difference did not meet the .05 threshold, t(73)=-1.92, p=.059.

Table 4. T-Test Results - Theoretical Concept Learning
Independent Samples T-Test Results of Theoretical Concept Learning by Participants

Theoretical Concept Learning	Mean	N	Std. Deviation	t	df	Sig.
R (open-source)	67.26	32	16.64	-1.92	72.96	.059
XLMiner (commercially-licensed)	76.05	43	23.01			

The third research question proposed was, "Is there a statistically-significant difference between an open-source data mining tool (i.e., R) and a commercially-licensed data mining tool (i.e., XLMiner), in terms of *practical skills learning?"* In order to address this research question, only the results from the *practical skills* test questions were analyzed.

Table 5: T-Test Results - Practical Skills Learning shows the comparison of average scores for both the R Programming class sections (M=64.45, SD=16.63) and the XLMiner class sections (M=66.96, SD=22.20) from the practical skills component of the assessment. The resulting t-statistic shows a *very slight* difference between the two software tool results, t(73)=-0.56, p=.575. In addition, the difference in learning for the practical skills questions was much smaller (t=-0.56), than the difference in learning for theoretical concept questions (t=-1.92).

Table 5. T-Test Results - Practical Skills Learning
Independent Samples T-Test Results of Practical Skills Learning by Participants

Practical Skills Learning	Mean	N	Std. Deviation	t	df	Sig.
R (open-source)	64.45	32	16.63	-0.56	72.99	.575
XLMiner (commercially-licensed)	66.98	43	22.20			

The fourth and final research question sought to determine which type of learning in this study was impacted the most by the choice of data mining tool. Specifically, the question asked, "Out of the three learning types (i.e., *Theoretical Concept, Practical Skills,* or *Overall*), which learning type is impacted the most by the choice of data mining tool?*"* In comparing the mean results from both software tools, the mean scores from the XLMiner sections were *slightly* higher than the R Programming results for overall learning, theoretical concept learning, and practical skills learning. For *overall learning*, the XLMiner mean (M=71.51) was slightly higher than the R Programming mean (M=65.86). For *theoretical concept learning*, the XLMiner mean (M=76.05) was slightly higher than the R Programming mean (M=67.26). For *practical skills learning*, the XLMiner mean (M=66.98) was also slightly higher than the R Programming mean (M=64.45). In order to determine if any of these three differences was at a statistically-significant level, the t-statistic was compared among the *overall learning* results, the *theoretical concept* results, and the *practical skills* results for both software tools. In comparing the results from **Tables 4 and 5**, the relative difference between the two tools was slightly higher for *theoretical concept learning* (t=-1.92), when compared to *practical skills learning* (t=-0.56), and to *overall learning* (t=-1.40). This may indicate (at least in the current study) that the use of XLMiner data mining software was *slightly* more effective than the R Programming Language at conveying the conceptual, theoretical components in the course material. While the theoretical concept learning difference in effectiveness

between the two tools approached statistical significance (p=.059), it should be noted that this difference in learning effectiveness between the tools did not meet the .05 threshold for statistical significance.

CONCLUSIONS

The current study sought to determine if the type of data mining software (i.e., open-source or commercially-licensed) had an impact on learning outcomes in an *Introduction to Data Mining* course. Specifically, the study compared a commercially-licensed data mining package (i.e., XLMiner from Frontline Solvers) to an open-source data mining package (i.e., R Programming Language). One of the major concerns in undertaking the study was the steep learning curve associated with the R Language. More precisely, the authors of the current study wondered if students would be able to handle both the theoretical aspects of machine-learning, and the detailed syntax of the R Language. Further complicating the class, the students were required to develop and evaluate their own data mining models within a short, eight-week course. For *overall learning*, the analysis in the current study revealed no statistically-significant difference between the use of a commercially-licensed data mining tool (i.e., XLMiner) and an open-source data mining tool (i.e., R). Further, the current study dove deeper into learning outcomes by examining two underlying components of student learning: *theoretical concept learning*, and *practical skills learning*. Again, the current analyses revealed no statistically-significant differences in student learning in neither the *theoretical concepts learning*, nor the *practical skills learning*. These findings suggest that the choice of software tool (i.e., open-source or commercially-licensed) has no meaningful impact on the knowledge transfer of machine-learning theory, nor the application of those theoretical concepts. In summary, either a commercially-licensed, or an open-source data mining tool should allow undergraduate and graduate students to apply theoretical concepts, develop and evaluate data mining models, and ultimately, apply machine-learning algorithms to solve "real-world" data problems.

The results of the current study are consistent with past studies that compared various types of data mining tools. For example, Zhang and Segall (2010) evaluated a combination of both commercially-licensed and open-source data mining tools. While the Zhang and Segall study ranked *SAS Enterprise Miner* as very high, the authors also gave equal standing to an open-source tool (*Megaputer PolyAnalyst*). Specifically, the authors of the 2010 study found that *both* SAS Enterprise Miner and Megaputer PolyAnalyst " . . . offer the greatest diversification of data mining algorithms" (Zhang and Segall, 2010, p. 652).

Other studies also ranked open-source tools as being "on par" with their commercially-licensed counterparts. Rangra and Bansal (2014) evaluated numerous data mining tools and ranked two open-source packages at the top of their list: KNIME and Weka. As an open-source tool, Weka was ranked as highest (or very high) in several studies (Wahbeh, Al-Radaideh, Al-Kabi & Al-Shawakfa, 2011). While Weka received high ranking in many of the studies, several other authors ranked R Programming among their top open-source tools and found it to be very effective for all data mining applications (Barlas, Lanning, and Heavey, 2015; Samad & Hassan, 2017).

While the current study was specific to XLMiner and R, the findings have broader implications for higher-education faculty who are considering various data mining tools for use in the Information Systems curriculum. Although the focus of the applied component in a course is choosing the best technology for improving learning outcomes, the choice of commercially-licensed versus open-source software may have broad implications for providing similar knowledge transfer across all Information Systems curricula. Of course, another important consideration is the choice of data mining software that best prepares students for the expectations of employers (e.g., the software tools used most in industry, and most frequently listed in job postings). In light of this consideration, the R Language is one of the most widely-used software tools, according to an annual poll by *KDnuggets* (Piatetsky, 2018). Therefore, the results of the current study are not only noteworthy in terms of improving learning outcomes, but also in terms of providing specific, marketable, data mining software skills to undergraduate and graduate students.

While the findings of the current study support the use of open-source data mining tools (such as R) in the Information Systems curriculum, further research is needed to maximize the learning outcome effectiveness of such tools. Future studies could address potentially confounding factors, such as different instructors using XLMiner and R. Such a study should involve the same instructor using R and XLMiner in different sections of the course, in an effort to eliminate any differences in teaching style that may influence the results.

In addition, the current study used a 20-question instrument, with closed-ended questions to gauge student learning with the two data mining software tools. A more robust evaluation of the data mining tools could include a more comprehensive test, with more questions (and various types of questions) to provide a more thorough assessment of student learning.

Finally, the current study focused solely on *learning outcome effectiveness* associated with the software tools. Future studies of data mining software tools could explore other criteria that are relevant to academia, such as ease of implementation, ease of use, and/or availability of teaching resources.

REFERENCES

Aasheim, C. L., Williams, S., Rutner, P., & Gardiner, A. (2015). Data analytics vs. data science: A study of similarities and differences in undergraduate programs based on course descriptions. *Journal of Information Systems Education*, *26*(2), 103-115.

Barlas, P., Lanning, I., & Heavey, C. (2015). A survey of open source data science tools. *International Journal of Intelligent Computing and Cybernetics*, *8*(3), 232-261.

Haughton, D., Deichmann, J., Eshghi, A., Sayek, S., Teebagy, N. & Topi, H. (2003). A review of software packages for data mining. *The American Statistician, 57*(4), 290-309.

Naur, P. (1974). Concise survey of computer methods. Petrocelli Books.

Piatetsky, G. (2018, May). 19th Annual KDnuggets software poll: Top analytics, data science, machine learning tools. Retrieved from https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

Provost, F. & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, *1*(1), 51-59.

Pushpalatha, C., Saravanan, V., & Ranjithkumar, C. (2014). Data mining open source tools - review. *International Journal of Advanced Research in Computer Science*, 5(6).

Rangra, K. & Bansal, K. L. (2014). Comparative study of data mining tools. *International Journal of Advanced Research in Computer Science and Software Engineering*, *4*(6), 216-223.

Samad, A., & Hassan, S. I. (2017). Assessment of the prominent open source and free data mining tools. *International Journal of Advanced Research in Computer Science, 8*(5), 2058-2062.

Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., & Al-Shawakfa, E. M. (2011). A comparison study between [sic] data mining tools over some classification methods. *International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*, 18-26.

Wang, J., Hu, X., Hollister, K., & Zhu, D. (2008). A comparison and scenario analysis of leading data mining software. *International Journal of Knowledge Management*, *4*(2), 17-34.

Zhang, Q., & Segall, R. S. (2010). Review of data, text and web mining software. *Kybernetes, 39*(4), 625-655.