DOES MACHINE LEARNING PROVIDE NEW INSIGHT FROM DATA: A COMPARISON OF CONVENTIONAL AND MACHINE LEARNING APPROACHES TO EXPLORE A DATASET

Sathiadev Mahesh, Department of Management & Marketing, University of New Orleans, New Orleans, LA 70148, USA, smahesh@uno.edu

ABSTRACT

Machine Learning (ML) has been used to analyze big data and provide targeted recommendations. ML toolsets such as Microsoft's Azure ML platform have made ML available to individual researchers. Can a researcher use these toolsets to obtain new insight from data? This paper explores the value of ML with smaller datasets and compares the results obtained by traditional statistical analysis on a dataset with ML analytics to look for new insights into the data. The slides are available at https://youtu.be/3ChBDYfhSKE.

INTRODUCTION

Machine Learning uses automated tools to analyze data and seek out patterns within the data set. In conventional statistical analysis the researcher builds a strong argument for a pattern within the data based on established theory, and then tests the data to verify support for the specified pattern. While it was possible for some researchers to violate the strict tenets of this approach and use powerful computer routines to run statistical analysis on data and seek out a statistically significant result, which could then be explained using theory, this approach was not always helpful since explainable and significant results were rare in small data sets. However, the availability of vast troves of data gathered by automatic surveillance tools, has allowed machine learning to throw up enough significant results. While ML cannot make theory and merely seeks out patterns, the results are useful for a business as long as these patterns are actionable. Predictive models need not start from theory; in fact, they can be data driven and remain useful until they are updated to reflect new data [6]. This paper uses ML to re-analyze data collected as part of global survey with over 1300 usable responses. The data which was analyzed earlier using conventional statistics to test a theory which classified the subjects in one of two categories, is re-processed using machine learning tools available on the Microsoft Azure ML platform. The results are compared with that from conventional statistics to test if (i) the previously found results are still supported, (ii) if new patterns are churned up by ML, and (iii) if they are usable.

Machine learning and analytics of big data are used by the most profitable organizations five times more often than by low performing organizations [10]. There are many ML models available for classification and the most popular among them are Support Vector Machines, Bayesian classification, Artificial Neural Networks, and Decision Forest Classification. This study tests these methods against traditional logistic regression. Logistic regression is used since the data has many binary variables.

PRIOR RESEARCH

A very early study on machine learning and the use of knowledge engineers working with human experts to generate rules for a DSS showed that a good induction based system that started with pre-defined and well-known rules could outperform a fully human generated set of rules [12]. Induction based systems are an early machine learning approach which used data from past business decisions to find rules that best matched the real-world data. Induction reduced the need for knowledge engineers to extract data from human experts and allowed records maintained by human experts to be "mined" for rules. One particular problem with the use of knowledge engineers was that they needed to thoroughly understand the field of expertise being modeled and there were relatively few people with this expertise. The human experts themselves were often unable to come up with usable rules for their expertise making rule development, and rule maintenance, the stumbling block for expert systems. These early induction systems however were stymied by the weak text analysis tools available at that time. Improvements in text data analysis have enabled ML tools to extract rules from data more effectively. Machine learning has been used to classify text data from online consumer reviews by form, function and behavior [16] demonstrating the improved capability of ML tools.

A comparison of the use of support vector machines, neural networks, and decision trees with multiple linear regression when discovering the underlying relationship between predictor factors and the usability of an e-Learning system shows that ML tools can help better identify usability factors for e-Learning systems [13]. Electronic medical records of Type II diabetes patients are analyzed using ML classification techniques to determine conditions under which certain medical decisions lead to sub-optimal results, and used to recommend alternative approaches earlier in the treatment cycle [11]. This study also shows that a similar approach can be effective when used on production shop floor data. A review of the use of ML in urban water management over the past decade shows ANN's, Bayesian networks, and swarm intelligence being used for supporting urban waste water decision support systems [7]. However, the same study also finds no evaluation of the relative merit of using ML over traditional case-based reasoning.

A time-series analysis used to forecast US aggregate retail sales shows that ANNs provide a better forecast (using MAPE) than conventional statistical methods [2]. Another, more recent study of different ML approaches on time series data uses the M3 competition data which contains over three thousand business time series datasets. This study showed differences in performance between ML approaches, with a relatively consistent ranking of the different approaches across different types of data [1]. However, this study also found that pre-processing of the data by conducting transformations and other analytics had a major impact on the quality of the results. Another study showed that pre-processing data by oversampling rare cases improved the effectiveness of ML approaches when the outcome being modeled is very rare [15].

When there is a small probability of a condition occurring in the data, such as in the case of financial fraud analysis from a mass of corporate reports, a traditional statistical approach, such as logistic regression performs very effectively and ANNs do not do as well [14]. Earlier studies on financial fraud used data with a higher likelihood of fraud and showed effective classification with ANNs. This shows the need to understand the behavior of the prediction variable before selecting a ML approach.

While powerful ML tools with user-friendly GUI interfaces enable business analysts and researchers to quickly and relatively painlessly try out ML, they have often been used without much concern for the validity of the underlying model. Many ML studies run the data on different ML models, a task that is

simplified by the new tools, and then report on the most effective model. This section reports many of those studies. However, few researchers have explored why a particular model work best for their problem. In the absence of a strong proof that a particular tool will continue to provide the best results for a class of problem, users will continue to run their data on many different tools and use the tool which provides the best result. ML makes it easy for the user to operate the model, while monitoring performance, until the model is no longer the best, and allow for a switch to "better" model at that time. This is similar to the framework of adaptive exponential smoothing in time series forecasting [18].

DATA ANALYSIS

Data collected from 1385 respondents was analyzed using conventional statistics (ANOVA, Factorial ANOVA, and Multiple Linear Regression). Rather than look at the specific instrument, factors, or coefficients of the data, this paper will merely study the outcome of the data analysis using conventional analysis. The same data was passed through Microsoft's Azure ML platform and tested using classification approaches.

As seen earlier in this paper, pre-processing data has been shown to have a major impact on result quality [1], [15]. The Azure ML framework incorporates many tools for easy pre-processing. In order to eliminate this effect, this study applied the same pre-processing to both approaches, logistic regression and ML tools from the Azure ML toolset. It should be noted that some of the ideas for pre-processing arose from a first round of ML analysis. This symbiotic learning process where the researcher/data analyst interacts with user-friendly ML tools and improves analysis, is the subject of a different paper. The pre-processing grouped the data in bins and selected randomly from the bins to neutralize differing counts for some of the underlying factors. The classification accuracy of the models is shown in Table I below.

Table I: Classification Accuracy of Models

Model	Precision/Accuracy	Comments
Logistic Regression – Conventional Statistics	0.668/0.733	Best performance on test data set.
Decision Forest ML	0.653/0.711	Provided many useful insights
Neural Network ML	0.638/0.707	Black box, no further insight into classification
Support Vector ML	0.662/0.726	Best ML model for this data

The next phase of this report reviews the usefulness of the model. The logistic regression and Support Vector models provided classifier weights shown in Table II. The Logistic regression weights are logit values and cannot be directly used in the manner of regression weights. They need to be converted into probability values. However, we can draw conclusions about the impact of factors on the odds of a subject being in Category I or II. Factors A through J are binary values (true/false). Hence, we see that when Factor C is true, it reduces the odds of a subject being in Category II, while Factors A and B strongly increase the odds of a subject being in Category II.

Logistic Regression fits the data to a single model. The result in Table II states that the probability of being in Category II (subjects can be in either Category I or Category II) can be determined from the weights.

Table II: Weights from Regression and Vector Model

Factors	Logistic Regression Weights	Vector Model Weights
Constant	-2.01	-2.26
A	2.08	4.79
B	1.68	3.76
C	-0.67	-2.97
D	.39	.34
E	.30	1.03
F	.19	.68
G	.16	.68
H	-.15	-2.74
I	.10	.68
J		-.68

The Decision Forest ML provides many trees to represent the data. In real-world scenarios a single valued common decision tree will represent only part of the data, and will not capture the rich complexity of real-world differences. Random forests generate multiple simple classifications that do not overfit the data [4], [8]. When a two-dimensional drawing is rendered as a 3-D object, it is found that random forests provide the best overall performance [17]. Random Forests are shown to have the better performance than all other benchmarks in predicting the impact of trading costs on high-volume automated market trades [3]. The best predictors of student progress are determined using a random forest approach using data student achievement and use of virtual learning environment data from a university in the UK [9]. A study of the performance of random forest based rankings of variable importance shows that the widely used mean-decrease accuracy approach is less stable than the rarely used mean-decrease Gini importance measure [5]. Many researchers merely seek out the best classification model and do not delve deeper into the model.

Table III: Links Gleaned from Decision Forest

Factors	Tree 1(i)	Tree 1(iii)	Tree II (i)	Tree II (ii)	Tree III (i)	Tree IV(i)
A	N	Y	Y		N	
B	Y	Y	Y		Y	
C						
D		N		Y		F
E				Y		T
F						
G	F				T	
H						

I						T
J			N	Y		
Cat	I-74%	II -100%	II- 92%	I -93%	II-71%	I-100%

This decision forest yielded many trees which provided insight to the data. A subset of the results is shown in Table III. The subjects can be classified in one of two categories, labeled as I & II. The trees in Table III show six of the results that can be gleaned from the decision forest.

If Factor A is No and Factor B is Yes and Factor G is No, then Cat I, Probability 74%
If Factor A is Yes and Factor B is Yes and Factor D is No, then Cat II, Probability 100%
If Factor A is Yes and Factor B is Yes and Factor J is No, then Cat II, Probability 92%
If Factor D is Yes and Factor E is Yes and Factor J is Yes, then Cat I, Probability 93%
If Factor A is No and Factor B is Yes and Factor G is Yes, then Cat II, Probability 71%
If Factor D is Yes and Factor E is Yes and Factor I is Yes, then Cat I, Probability 100%

Factor J is not present in the logistic regression but has an impact on two of the rule extracted from the decision forest ML tool. In a similar manner many other conclusions can be drawn from the trees in the decision forest. This yields many more rules for interpreting the data. It is possible to obtain rules that are not fully consistent, but have partial support for a conclusion. Machine Learning opens up new classification approaches that can induce rules from data which are not clearly seen from conventional statistical analysis, especially when the models are linearized for simplicity. Much of the variety on real-world data is lost in the effort to fit an overarching, single statistical model, often linear, to the data. Machine Learning tools open up new analytic approaches.

SUMMARY AND CONCLUSIONS

This paper reports on the use of machine learning tools to seek patterns in data previously analyzed using conventional statistics. While the overall classification quality does not improve in this ML analysis, the tools offer a rich suite of options which reveal patterns within the dataset that can help better understand the interplay of factors in the data. Black box ML tools such as ANNs do not help in improving the understanding of patterns within the data. However, decision forest tools show promise for the discovery of many patterns within the data.

REFERENCES

[1] Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. An Empirical Comparison of Machine Learning Models for Time Series Forecasting. *Econometric Reviews*, 2010, 29(5/6), 594-621. doi:10.1080/07474938.2010.481556

[2] Alon, I., Qi, M., Sadowski, R. J. Forecasting aggregate retail sales: a comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services*, 2001, 8:147–156

[3] Booth, A., Gerding, E., & McGroarty, F. Performance-weighted ensembles of random forests for predicting price impact. *Quantitative Finance*, 2015, 15(11), 1823-1835. doi:10.1080/14697688.2014.983539

[4] Brieman,L, Random Forests, *Machine Learning*, 2001, 45(1), 5-32, doi: 10.1023/A:1010933404324

[5] Callean, M. L., & Urrea, V. Letter to the Editor: Stability of Random Forest importance measures. *Briefings in Bioinformatics*, 2011, 12(1), 86-89. doi:10.1093/bib/bbq011

[6] Dhar, V. Data Science and Prediction. *Communications of the ACM*, 2013, 56(12), 64-73. doi:10.1145/2500499

[7] Hadjimichael, A., Comas, J., & Corominas, L. Do machine learning methods used in data mining enhance the potential of decision support systems? A review for the urban water sector. *AI Communications*, 2016, *29*(6), 747-756. doi:10.3233/AIC-160714

[8] Ho, T.K. Random Decision Forests, *Proceedings of the Third International Conference on Document Analysis and Recognition, ICDAR '95*, August 14-15, 1995, 1, 278

[9] Hardman, J., Paucar-Caceres, A., & Fielding, A. Predicting Students' Progression in Higher Education by Using the Random Forest Algorithm. *Systems Research & Behavioral Science*, 2013, 30(2), 194-203. doi:10.1002/sres.2130

[10] LaValle, S., E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz. Big Data, analytics and the path from insights to value. *MIT Sloan Management Review*, 2011, 52 (2): 21–32

[11] Meyer, G., Adomavicius, G., Johnson, P. E., Elidrisi, M., Rush, W. A., Sperl-Hillen, J. M., & O'Connor, P. J. (2014). A Machine Learning Approach to Improving Dynamic Decision Making, *Information Systems Research*, 25(2), 239-263. doi:10.1287/isre.2014.0513

[12] Michaelsen, R. H., & Swigger, K. M. Analysis of the Effectiveness of Machine Learning in Determining Decision Rules for Executive Compensation Planning. *International Journal of Intelligent Systems In Accounting Finance & Management*, 1994, 3(4), 263-278.

[13] Oztekin, A., Delen, D., Turkyilmaz, A., & Zaim, S. A machine learning-based usability evaluation method for eLearning systems. *Decision Support Systems*, 2013, *5*,663-73. doi:10.1016/j.dss.2013.05.003

[14] Perols, J. Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *Auditing: A Journal of Practice & Theory*, 2011, 30(2), 19-50. doi:10.2308/ajpt-50009

[15] Perols, J. L., Bowen, R. M., Zimmermann, C., & Samba, B. Finding Needles in a Haystack: Using Data Analytics to Improve Fraud Prediction. *Accounting Review*, 2017, 92(2), 221-245. doi:10.2308/accr-51562

[16] Singh, A., & Tucker, C. S. A machine learning approach to product review disambiguation based on function, form and behavior classification. *Decision Support Systems*, 2017, 9,*7*81-91. doi:10.1016/j.dss.2017.03.007

[17] Soundararajan, K. P., & Schultz, T. Learning Probabilistic Transfer Functions: A Comparative Study of Classifiers. *Computer Graphics Forum*, 2015, 34(3), 111-120. doi:10.1111/cgf.12623

[18] Trigg, D.W. & Leach, A.C. Exponential Smoothing with an Adaptive Response Rate, *Operations Research Quarterly*, 1969, 18, 53-59.