Paper 1934-2018

Are New Modeling Techniques Worth It?

Tom Zougas, TransUnion

ABSTRACT

New modeling techniques are constantly being developed, and at an increasing pace. Keeping up with these techniques is possible if the technique is accompanied by a library or an example that implements the technique in one of the many analytic tools at our disposable. However, the main question will always be whether the new technique is adding value and actually addressing the business problem at hand. This paper addresses the question of when to apply new modeling techniques, and more importantly, when not to.

INTRODUCTION

Whether you are a seasoned data scientist, or in the early stages of becoming a data scientist, it is important to have a core set of models to use for analyzing data. These core types of models provide the basis for solving a business problem and effective use of available data. Keep in mind that the field of data science is constantly growing with availability of more and more data, and new techniques and models for analyzing the data, but you don't want to get sidetracked from focus of solving the business problem using data.

This paper will not be a survey of modeling techniques. There are existing papers that have already documented that, such as (Das and Behera, 2017; Wu et al 2009). This paper will focus on the strategy behind solving business problems by applying the right models given the available data.

KNOW THE CORE TYPES OF MODELS

In a recent KDNuggets survey (Piatetsky, 2017), the top 3 of the top 10 machine learning methods used were shown to be:

1. Regression
2. Clustering
3. Decision Trees[1]

This is the same as previous years' surveys.

What is interesting about these 3 techniques is that they are not what one would consider state-of-the art, or leading edge, or exotic. However, they have been around for a while and are fundamental, core, and what you could consider go-to models. What they do is provide a starting point for an analysis and a baseline result to measure any other technique against. Ultimately, any new technique can be measured against these core models to determine if the new technique would provide any additional value to solving the business problem.

The reasons why these 3 modeling techniques continue to be used, and why they rank as the top 3, can be attributed to the following:

- They are simple
- They are interpretable
- They work

[1] Visualization tied with Decision Trees, but for the purpose of this paper, we are not considering Visualization as a machine learning model

It is for these reasons that the data scientist should include and use these techniques as part of their basic toolkit during modeling. As it turns out, these core techniques are available within many software packages or languages used by the data scientist.

SUPERVISED AND UNSUPERVISED LEARNING

There are 2 main categories of algorithms that are used for solving a large number of business problems: supervised learning and unsupervised learning.

Supervised learning is applied when you have a dataset containing input attributes (variables, characteristics) and a known output (outcome, target) associated with each record or observation. The output can be discrete or continuous. The machine learning algorithm determines how the inputs relate to the output, providing a means of inferring the output for a new set of inputs.

Unsupervised learning is applied when the dataset does not contain a known output. The algorithm tries to identify hidden patterns in the data usually in the form of categories or classes.

With the 3 core modeling techniques, the data scientist has the ability to perform either supervised or unsupervised learning, where regression and decision trees can be applied for supervised learning and clustering can be applied for unsupervised learning.

There are several additional algorithms available for supervised and unsupervised learning, listed in Table 1. Each algorithm has its strengths and weaknesses. However, before looking into using these other types of models, the recommendation is to start with the core models identified above to provide a baseline. Often times, the baseline is sufficient to satisfy the business objective.

Category	**Algorithm**
Supervised Learning	Support Vector Machines
	Linear Regression
	Logistic Regression
	Naive Bayes
	Linear Discriminant Analysis
	Decision Trees
	K Nearest Neighbor
	Neural Networks
Unsupervised Learning	Clustering (k-means, hierarchical)
	Anomaly Detection
	Neural Networks (autoencoders, SOM)
	Expectation–Maximization
	Principal Component Analysis
	Singular Value Decomposition
	Association Analysis

Table 1. List of Supervised and Unsupervised Learning Algorithms

APPLYING THE CORE MODELS

So where would a data scientist use the core models? This depends on a couple of factors: what is the business problem you are trying to solve, and what is the form of the data you have available. These are the key starting points associated with the overall modeling process, which we describe below.

It is widely accepted that the modeling activity typically takes 10-20% of the overall effort associated with building models. The remaining 80-90% of the effort is spent in defining the objective, exploring and manipulating the data, and finally deploying the model. This is concisely described by the CRISP-DM methodology (CRISP-DM) and illustrated in Figure 1.

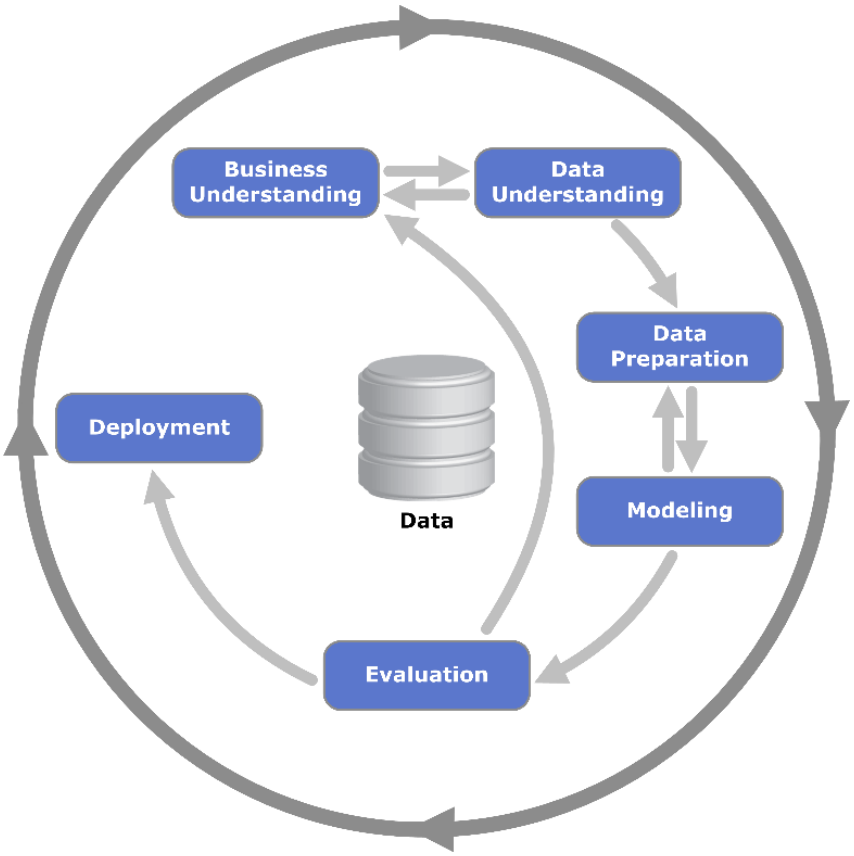

Figure 1. The CRISP-DM Methodology[2]

Following the CRISP-DM methodology as the basis, the following highlights some of the key considerations in solving a business problem using machine learning.

WHAT IS THE BUSINESS PROBLEM

This is the starting point of any analytics exercise: defining the business problem. For example:

- Do you want to categorize a record or observation?
- Do you want to predict a numerical quantity?
- Do you want to rank order your records based on some outcome of interest?

[2] By Kenneth Jensen - Own work based on: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf (Figure 1), CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24930610

- Do you want to identify naturally occurring groupings or clusters in the data?

The importance of defining the business problem cannot be understated. It is essential for ensuring you obtain the right data, and identifying which model type to apply. In addition, it provides a measure for determining if you have been successful with the analysis, and when you are done and ready to deploy to production.

DATA EXPLORATION AND PREPARATION

This is where most of the effort is typically spent: working with the data. You have to collect it, usually from multiple sources; manipulate it to get it into a usable form; explore it to see if it is useful, and of good quality; and ultimately determine if it will support the business objective. If there are issues with the data, then decisions have to be made in terms of adjusting the business objective, pulling more data, or deciding the business problem cannot be solved.

If the data is usable, then you will move onto the modeling process.

THE MODELING PROCESS

The modeling process consists 2 steps as described by the CRISP-DM methodology: modeling and evaluation. The modeling step is where you apply one of the relevant (and in this case, core model types). The selection of model types is rather straightforward:

- If the data contains known outcomes, then you use a supervised learning algorithm: regression or decision tree (or possibly both).
- If the data does not have known outcomes (or labels), then apply the unsupervised learning algorithm: clustering.

Given the ease and speed with which the 2 supervised learning algorithms can be applied, it is common to build both models and then compare the results, selecting the one that works the best.

The evaluation step is where you determine if the model is actually working for the business problem at hand, and whether improvements are required. The type of improvements will depend on whether the issue is with the data or with the model's ability to extract information from the data. If it is the latter, then that would be a valid reason to explore more sophisticated model types that can better capture the patterns within the data. If the issue is with the data, then there is no reason to try other model types, but better to focus on improving the content and quality of data being used.

There are various tools available to do the model evaluation and is dependent on the type of outcome. For example, if you are modeling a binary outcome, the Receiver Operating Characteristic (ROC) chart is a powerful way to assess the performance of the model, as demonstrated in Figure 2. For this example ROC, the Random line illustrates no lift from the model (the model is no better than random selection); models 1, 2, 3 provide increasing lift, respectively. It is up to the data scientist to determine how much lift is sufficient to satisfy the business objective, and how much additional effort is worth pursuing to go from 1 to 2 and from 2 to 3.

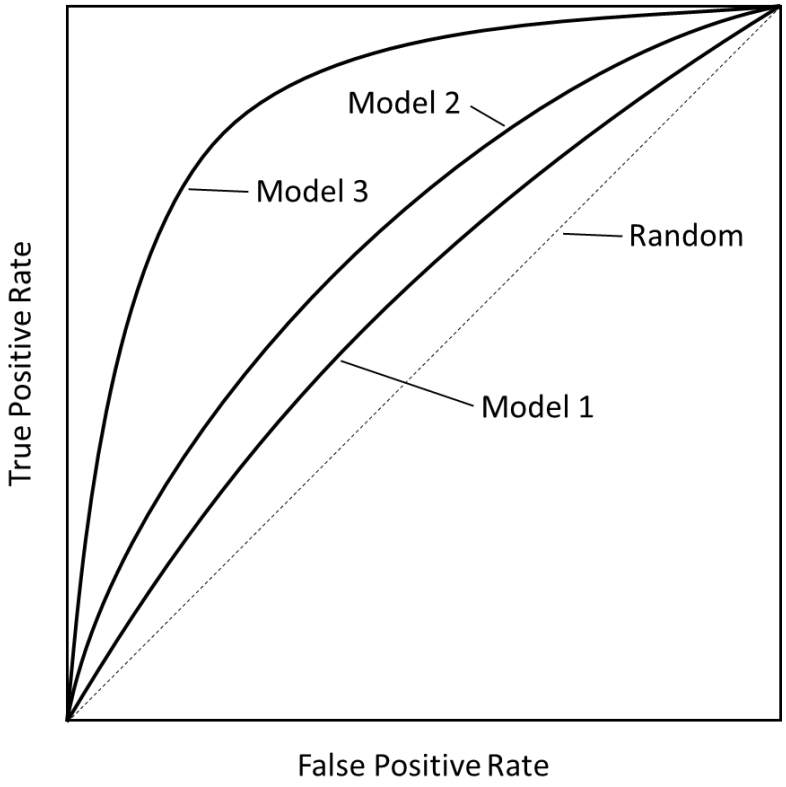

Figure 2. ROC Chart for Assessment Model Performance

Evaluating the results from unsupervised learning is more difficult in that there is no known outcome to compare with. For unsupervised learning, evaluation is up to the data scientist to determine whether the clusters produced by the model are of value by looking at the profile of the attributes within the clusters to understand what the clusters represent.

Ultimately, a decision has to be made whether the model results have satisfied the business objective, or if it is worth trying something beyond the core models to determine if the results can be improved upon or not. Often times, the core models provide sufficient value especially when you consider that they are simple, interpretable, and they work.

MAKING USE OF THE MODEL

Once the model is complete, you have to decide how it will be used. Is the intention to simply report on the investigation and provide the insights gained from the model, or is the need to deploy the model into a production environment to leverage the model's predictions on new data?

This step in the process is important in that it closes the loop on the initial problem definition. It confirms whether the model selected can effectively be used in production, and the effort required to deploy the model. Using one of the core models facilitates ease in deploying the model since the coding tends to be straightforward.

A SAMPLING OF NEW MODELS

If the core models used do not work, a decision has to be made as to whether the problem is with the data available, or the ability of the model to use the data effectively. Let's assume the data exploration has indicated there is value in the data, but somehow the core models are just not able to pick out the necessary patterns, or the lift obtained is barely present or of insufficient value. This is the point where trying new models is warranted. Keep in mind that the context of what we are describing here is using

models in a commercial environment; if the purpose of the work is more research oriented, then working with some of the newer and more sophisticated models may be more appropriate.

The following is a sampling of what new models exist. For a broader view, refer to (Pham, 2017).

DEEP LEARNING

Deep learning is fundamentally Artificial Neural Networks (ANN), which have been around for a while. What has made deep learning popular in recent years is the availability of large volumes of data (think Big Data), and substantial computing power to consume and process the data.

The earlier implementations of ANN worked in solving similar problems as the above core models were able to solve. The ANN would provide models which worked as well as, and sometimes better than, regression or decision trees. This would usually be attributed to their ability to represent a wide range of nonlinear patterns in the data, which regression and decision trees might struggle to represent. However, their reduced popularity would be attributed to an inability to explain the results. They worked, but they were not easy to understand. This lack of interpretability is something that many businesses would not be able to accept; they want to know what is it about the inputs that results in the specific predictions produced by the model.

However, there are use cases such as Fraud Detection where the desire to have a model that works really well is more important than being able to explain the reason for the results. Additional use cases where deep learning has been successfully applied more recently include (Deep Learning):

- Automatic speech recognition
- Image recognition
- Visual art processing
- Natural language processing
- Drug discovery and toxicology
- Customer relationship management
- Recommendation systems
- Bioinformatics
- Mobile advertising
- Image restoration

ENSEMBLE

There are different approaches to ensemble modeling, from using random forest, to boosted trees, to voting from several base models. The general idea of ensemble modeling is that if one model can find some patterns, then an army of models can probably find more patterns, which when combined can provide a stronger model.

Ensemble models have been effectively used to win Data Science competitions, such as Kaggle[3]. But competitions aren't the same as successfully solving and deploying a model in a business or commercial environment. What winning competitions does is demonstrate that it is possible to get an edge above most models by using ensembles. So if that extra bit of lift has significant business value, then it is worth trying. But be aware that there will be more effort needed to build the model, and depending on how you plan on deploying the model, the effort to deploy may be expensive.

When should the data scientist apply ensemble models?

- When a single algorithm does not approximate the true prediction function well.

[3] www.kaggle.com

- When model performance is the most important factor (over training speed and interpretability).

One specific type of ensemble model which has been gaining popularity recently is XGBoost (XGBoost), which is an implementation of gradient boosted decision trees designed for speed and performance. It has been implemented in multiple languages including C++, Python, R, Java, Scala, Julia so can be used by data scientists regardless of what language they use. XGBoost may become one of the new core models if it stands the test of time and delivers on its objectives of speed and performance. In addition, efforts have been made to provide interpretability to XGBoost models (Foster, 2017).

ELASTIC NET

Elastic Net provides a means of improving on regression models, especially in the case of high dimensional data. It reduces the dimensionality of the data by selecting a subset of variables. It works towards grouping related variables in the selection process. In addition, it is able to handle variables which are highly correlated. The result is a model with improved accuracy and interpretability.

It is available in several languages such as SAS, R, Python and Spark.

AUTOMATED MACHINE LEARNING

There are many new models that have become available in recent years. The challenge is knowing which models to use as it is difficult for the data scientist to know how to apply each of the new models. This is where Automated Machine Learning (AML) comes in. The goal of AML is to sift through a wide range of modeling algorithms and based on one or two performance criteria, determine the best model. Where ensemble models try to combine the results of several models, AML tries to determine out of a large library of models, which one is the best one to use. It is like having a master data scientist perform all the work of training and testing possibly hundreds of models, but it does it much faster than the data scientist could.

It may appear that AML will result in no longer requiring a data scientist to be involved. On the contrary, what it does is automate the mundane tasks that used to be performed by the data scientist, similar to how spreadsheets automated the mundane tasks that accountants used to do.

Although AML may be successful by providing improved accuracy and speed in building models, there is a tendency for them to become less interpretable (more of a black box), which again may not be suitable for all business environments.

CONCLUSION

To summarize, the recommendation is to start with the core models. They provide a good starting point for solving a business problem. If the business problem is similar to one that has been solved before, you would have an existing process to follow in terms of collecting and manipulating the data, building the model and then deploying it. And if the result is sufficient for the business objective, then you have been successful.

If the core model types are not able to satisfy the business objective, then ask why: is it a problem with the data or a problem with the selected model being able to effectively extract the information from the data? If the problem is with the data, then a new model won't fix bad data. Also, if the new model is overly complex, and difficult to build and deploy, then make sure there is sufficient added value in the results of the model to warrant its use. There are many modeling techniques available; you don't want to get lost in trying so many that you lose sight of solving the business problem at hand.

REFERENCES

Das, Kajaree and Behera, Rabi Narayah. 2017. "A Survey on Machine Learning: Concept, Algorithms and Applications." *International Journal of Innovative Research in Computer and Communication Engineering,* 5(2):1301-1309. Available at http://www.rroij.com/open-access/a-survey-on-machine-learning-conceptalgorithms-and-applications-.pdf

Wu, Xindong, Kumar, Vipin, Quinlan, J. Ross, Ghosh, Joydeep, Yang, Qiang, Motoda, Hiroshi, McLachlan, Geoffrey J., Ng, Angus, Liu, Bing, Yu, Philip S., Zhou, Zhi-Hua, Steinbach, Michael, Hand, David J. and Steinberg, Dan. 2008. "Top 10 Algorithms in Data Mining." *Knowledge and Information Systems*, 14:1-37. Available at http://www.cs.uvm.edu/~icdm/algorithms/10Algorithms-08.pdf

Piatetsky, Gregory. 2017. "Top Data Science and Machine Learning Methods Used in 2017." Available at https://www.kdnuggets.com/2017/12/top-data-science-machine-learning-methods.html

CRISP-DM. In *Wikipedia*. Retrieved March 6, 2018, from https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Deep Learning. In *Wikipedia*. Retrieved March 6, 2018, from https://en.wikipedia.org/wiki/Deep_learning

Pham, Thuy T. "Top 20 Recent Research Papers on Machine Learning and Deep Learning." 2017. Available at https://www.kdnuggets.com/2017/04/top-20-papers-machine-learning.html

XGBoost. "Scalable and Flexible Gradient Boosting." Available at http://xgboost.readthedocs.io/en/latest/

Foster, David. 2017. "NEW R package that makes XGBoost interpretable." Available at https://medium.com/applied-data-science/new-r-package-the-xgboost-explainer-51dd7d1aa211

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Tom Zougas
TransUnion
tom.zougas@transunion.com