Model Selection for Machine Learning Algorithm on Decision Making in Oil and Gas Upstream Project Malaysia

Mohd Shahrizan Abd Rahman¹ and Nor Azliana Akmal Jamaludin²

¹Petronas, Level 19, Tower 3, KLCC 50088 Kuala Lumpur
²Faculty of Defence Science and Technology, Universiti Pertahanan Nasional Malaysia, 57000 Kuala Lumpur
Email: mshahrizan.ar@petronas.com.my

Abstract. Model selection is a crucial element in data analysis to get reliable and reproducible statistical inferences or predictions. It is a long history of model selection method arising from research in statistics, information theory, and signal processing. The purpose of this study is to address the problems related to big data in contributing the strategies to make decisions on new investments for upstream Oil and Gas projects in Malaysia. It also discusses the use of machine learning methods for big data processing and highlights current scenarios in a model selection perspective. Machine learning algorithms have proven to work well for statistics used to make decisions. The selection of the machine learning algorithm model does not make drastic assumptions about data, and it can help optimise the exploration process and allow the computer to analyse large amounts of data quickly and accurately. The results show that k-fold cross-validation of the developed model options intended to make subsequent decisions because it is an integral portion of big data processing to gather unexpected new insights, discover new knowledge and improve efficiency.

1. Introduction

Strategic decision making on valuable resources should measure to avoid risk. Many areas of the practical application of risk as regards the decision-making center. However, machine learning is not; this is because research should provide insights and algorithms that provide machine learning with the ability to consider theory-decision risks [1]. Significant advances in data mining techniques are robust and scalable, fast method to detect large databases, machine learning, and innovative applications for business applications has come from IBM research labs worldwide [2].

The challenge of machine learning occurs when choosing between models’ selection that can be used operationally for months or years. In this study, we will find model selection challenges for machine learning problems when modelling beyond model performance, such as complexity, maintenance, and available resources. The model selection in machine learning applied to understand user behaviour, goal advertising better, expand functioning efficiency, reduce expenses, and decrease risk. International Data Corporation (IDC), is a provider of global information technology and market intelligence services, predicts that the big global data and analytics market will grow in the future [3].

1.1 Machine Learning Meaning

Machine Learning is an Artificial Intelligence field related to the use of computers to simulate human learning, which enables computers to identify and acquire knowledge from the real world and enhance
their performance in some new knowledge-based tasks. Today, Machine Learning algorithms benefitted in several areas not limited to computer science, such as business [4], advertising [5], and medicine [6].

![Figure 1. Workflow for Machine Learning Project by David Chappel](image)

The objective machine learning is to analyze algorithms in which the system will categorize problems and solve problems without human assistance or supervision. This system will demonstrate the ability to identify, learn, transform, develop and run independently when new data presented. Machine Learning focuses on the possibility of developing a system to obtain information and use it alone [7]. In the growing demand for smart applications driven by business data, such as IBM companies, are developing a large number of applications that use machine learning technology to optimize business process decisions [8]. The machine learning concept extensively in business applications developed by IBM for internal and customer use. As a machine learning application used in operating IT environment, scalability requirements, durability, and automatically attract attention.

It is because, the in-situ learning needs of extensive data have run to the need for the design and implementation of scaling machine learning algorithms, and a framework for parallel machine learning technologically advanced to enable approaches based on standard development algorithms. Therefore, it can leverage specialized hardware accelerators and new multi-core architecture in the business world. Besides, applications with a high-volume low latency flow environment drive real-time and online machine learning methods. The relationship between machine learning and optimization takes on a whole new dimension and also, the role of optimization methods in machine learning of modern algorithms needs to be well understood [9]. However, when develop a decision support system for end-to-end services that combine predictions and prescriptions. The system is a growing need for a combination of predictive modelling and optimization solution to produce a plan.

1.2 Machine Learning Methods
Interest in complex diagnostic instrumentation systems has also driven new directions in machine learning research. The earliest application of machine learning started through the marketing area and relationship management. The application developed and fully utilize for management observation, targeted marketing and cross-selling advice—example of methods such as decision making, rule induction, and collaborative filtering. Machine Learning algorithms alienated into supervision, supervisory and non-supervisory partial. In supervised learning, input, and the desired output set out, and the algorithm used to study the function of mapping from input to output [10]. In supervised learning, two main tasks are identified, such as classification and regression. Classification can mean as the output that predicts the target class while in the regression output is to predict the continuous value. K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive-Bayes (NB) are algorithms for classification [11]. Algorithms such as neural networks, Naive-Bayes and Support Vector Machine can be used for this purpose. Partially controlled culture preferred only when the removed labelled data requires a skilled and relevant source to learn from it.
Table 1. Summary of Several Machine Learning Algorithms

| ALGORITHMS                  | TYPES            | CHARACTERISTICS                        | LEARNING POLICY                        | LEARNING ALGORITHMS          | CLASSIFICATION STRATEGY          |
|-----------------------------|------------------|----------------------------------------|----------------------------------------|-------------------------------|----------------------------------|
| Decision Tree               | Discriminant     | Classification tree                    | Regularized maximum likelihood estimation | Feature selection, generation, prune | IF-THEN rule according to tree spitting |
| Non-Linear SVM (based on libsvm) | Discriminant     | Super-plane separation, kernel trick | Minimizing regular hinge loss, soft margin maximization | Sequential minimal optimization algorithm (SMO) | Maximum class of test samples |
| Linear SVM (based on liblinear) | Discriminant     | Super-plane separation                 | Minimizing the loss of regular hinge. Soft margin | Sequential dual method       | The maximum weighted test sample |

1.3 Machine Learning Category

Three primary categories identified in machine learning as per below explanation:

- **Supervised learning**: This is the procedure of drawing information from a set of observable outcomes. Classification of data mining is a technique that gives the object to one of several pre-defined categories [12].
- **Unsupervised Learning**: It expended to extract data from a set of information whose results are unknown. Also known as platform that classifies customers into several different profiles and lifestyles.
- **Reinforcement Learning**: This method is not accessible or used in general because its application is limited to small groups only. This type of learning is also called trial and error learning.

In the extremely modest world of today, the value and timeliness of business information for the organization is not just a choice between the benefits and burdens; it may be a inquiry of resilience or disaster. The rapidly changing business environment will increase machine learning needs as a business intelligence tool [13].

2. Model Selection

Model selection is the art of choosing from a variety of mathematical models that give a good idea of a particular real-world phenomenon. There are many reasons why a model selection may essential in a decision-making application. Some of the main reasons are that, for the first time, the inclusion of unimportant models adds unnecessary noise to the task of optimising its fundamentals and the inclusion of incorrect interactions can cause bad things to happen [14].

Oil and gas production company use thousands of sensors installed in underground and surface facilities to provide continuous data collection, monitoring real-time asset and environmental conditions. This data comes in a structured, semi-structured and non-structured form. According to Gupta [15],
analysis reveals the patterns and relationships of vessels in this data to improve decision making. The analysis technique used to identify patterns in historical data and even species that can then be associated with current or future data to identify risks and opportunities [16].

2.1 Method of Model Selection
For example, if the purpose of decision-makers is using the model to obtain qualitative insights or make predictions, this criterion is likely to perform well. However, these criteria encountered to choose a model that produces excellent results. Claesken, et al. [17], then proposed a method, Focused Information Criterion (FIC), which aims to select a model that provides an excellent accuracy to estimate the parameters of interest. The FIC’s idea is to estimate the average error of interest parameters for each model, and then select a model that has a minimum budget.

2.1.1. Statistical Learning Theory. Statistical learning theory Hastie et al. [18], addresses questions that are strictly related to model selection. The objective in this area is to build a functional prediction function \( f: x \to y \) which provides a good description of the relationship between the random input variable \( x \) with support \( x \) and the output of the random variable \( y \) with support \( y \). Joint distribution \( (x, y) \) is unknown, but data consisting of realisation using \( (x_i, y_i) \) \( 1 \leq i \leq n \) of \( (x, y) \) is available. As described in Bousquet et al. [19] and Guyon et al. [20], the primary method to determine good predictors namely empirical risk reduction, structural risk reduction, regularisation method based on the idea of minimising the empirical risk \( \sum_{i=1}^{n} L(y_i, f(x_i)) \) on some class \( G \subset y^x \) predictors may couple with the complexity of the model terms that penalise \( f \). Selecting \( G \) can be viewed as a model selection problem.

2.1.2. Bayesian Model Average. Bayesian Model Averaging (BMA) is a Bayesian application conclusion on model selection problems, combined assessments and predictions that result in simple model selection criteria and risk predictions. However, the implementation of BMA is not always easy, leading to a variety of assumptions and options related to different aspects [21]. The average Bayesian model [22], however, is an approach that can link optimisation problems with different models. The main idea is not to select a Bayesian model of an alternative model but to spread the distribution of the probability that any given model is correct, and this applies to all of the probability distribution of the next generation. In briefly, BMA marginalises the model to obtain posterior density on model parameters that contribute to model uncertainty, as follows:

\[
p(\theta | y) = \sum_{m_i} p(m_i | y) p(\theta | y, m_i)
\]

where \( m_i \) is the candidate model group, \( p(m_i | y) \) is the posterior probability of this model, and \( p(\theta | y, m_i) \) is the posterior density model parameters depending on the model \( m_i \). Last posterior density is a good proxy for information about the parameter \( \theta \) someone only if \( p(m_i | y) \approx 1 \). Otherwise, the uncertainty about the correct model will automatically lead to uncertainty about model parameters.

2.1.3. Facilitate Model Selection and Data-based Optimisation. The idea that the model selection methods must be consistent with the objectives they produce excellent results implied in some recent studies. Bastani et al. [23] considered the problem of lineaments armed robbers with high dimensional covariates and adapted the LASSO estimator parameters [24] to achieve optimal symmetric rewards. Because this parameter is a measure of the complexity of the regulatory model, their method can become an example of adjusting the model selection method in order to produce good results. A similar idea appeared in [25], which increases the portfolio optimisation problem with a data-driven differentiation parameter that limits the sample variance of the objective function is estimated. Parameter adjustment is optimised variant of K-fold cross-validation, where validation ignored by performance metrics that are relevant to the issue of investment.

2.2 The Algorithm Model Selection
Non-inclusive algorithms, to explain in detail the issues and challenges in advanced learning methods. Several methods have executed to develop machine learning algorithms in order to entree large-scale
data such as MapReduce and spread templates such as Hadoop. Advanced learning and the extraction of various stages of data acquisition within learning within provide an extraordinary level of explanation for Big Data Analysis [29]. Quality of decision-making enables businesses to adapt faster than their competitors to the opportunities and threats posed by the development of the digital age and the market [30]. Machine-learning enables the computer to find the hidden views without programming clearly where to look with the advent of big data capable of automatically applying sophisticated mathematical calculations such as recent developments, coupled with increasing volume, processing and type of data, as well as affordable data storage, resulting in faster and smarter data usage for more accurate decision making [31]. Decision making always presented as a rational process, in which individuals make decisions by collecting, integrating and analysing data mechanically and rationally. Nevertheless, research has long shown that this is not how people make decisions. Furthermore, this is where deep learning become dominant, as a subset of machine learning related to algorithms inspired by brain structures and functions called artificial neural networks [32].

2.3 Predictive Analysis Machine Learning
Oil and gas leader must perform machine learning in addition to the engineering model right relating to the Industrial Internet of Things (IIOT). Some forms of machine learning predictive analytics can be given in oil and gas sector, including [33]: i. Predictive Maintenance - With a integration of machine learning and application maintenance to deliver IIOT data utilising equipment failures budget more accurately, achieving positive outcomes and reduce the cost, time stops, and the risk of the investment. ii. Reservoir Modelling - Machine learning makes the process and procedure more proven and faster decision making by providing data that is in the reservoir history matching recognise patterns. iii. Video Interpretation - As many sensors that are part of the process of continuous monitoring of platform or refinery, video technology used in down-hole drilling environments can utilise machine learning. iv. Reasoning based on case - Current problem with historical cases to find sameness that can come up with indication to give assistance to identify actions or behaviours that can help prevail over current situations. Including analysing data such as equipment used, depth, weather conditions, costs and more [34].

3. Result
The model selected using classical criteria relating to obtaining statistical estimates with a small (such as a mean squared error). However, as illustrated in Figure 2, a small statistic does not show that the model selected has good results. Practical examples are clear about this phenomenon [26], which compares advanced machine learning models with simple multinomial login (MNL) models, in the problem of optimising product display in large online markets.

![Figure 3](https://example.com/Figure3.png)

**Figure 3.** Example of Illustrations of Significant Results by Cooper, William L. Besides, Besbes et al. [27] and Cooper et al. [28]
It has been shown that incorrect models can sometimes lead to good or better results than ‘correct’ models. However, in the case of data-based optimisation, the value of the model should be evaluated based on the quality of the results it produces. Figure 1 shows that even the objective functions estimated by Model 1 are closer to the truth than Model 2 (measured, for example, with their L² distances), the optimal \( x (2) \) yields corresponding to Model 2 provide higher \( f (x (2)) \) of the results. The optimum \( x (1) \) corresponds to Model 1. In making decisions, the choice of model is usually either complex, ‘realistic’ and ‘simple’ or simple.

4. Discussion
K-Fold Cross Verification is where a set of data that is divided into several sections \( K \) / folding in which each fold sourced as a test set at a time. The main idea is to maintain an independent test data set, during training and model selection, to prevent leakage of test data at the training stage:

![Figure 4. K-fold Cross-Validation: Evaluate the Performance of the Estimator](image)

The process is straightforward and matching to the three-way workflow detection by dividing the data set into two parts, one training and a set of independent testing to measure end model evaluation. In the second step, the various settings hyperparameter with the use of Bayesian optimisation, Grid Search standard or Randomized Search. The researchers used k-times cross-validation on the training set in order to produce several models and performance estimates for each hyperparameter configuration. Using a full set of exercises for the assembly model and taking hyperparameter settings to match the best performing model [35]. Finally, after the evaluation is complete, the model can load all the data, which can be a deployment model. When it comes to in-depth learning literature, when facing to model evaluation, researchers often to choose the 3-way detention method because of practicality and efficiency besides also common in writing extended learning (not rooted). From the cost perspective and saving, retaining three directions is preferred over k-fold cross-validation. Apart from the problem of computing efficiency, using deep learning algorithm when the sample size is relatively large, a scenario in which we do not have to worry about the high variance because of the sensitivity of our estimates of how we divide the data set for training, proof, and challenging.

For the future, the researcher decided to choose the type of k-fold cross-validation one of model selection which can develop for machine learning algorithms on decision making in oil and gas upstream project Malaysia. Model selection of machine learning based on algorithms are an inseparable part of big data processing to gather unexpected new insights, discover new knowledge and improve efficiency. Nowadays, the method used for model selection in sequential decision making steered by expert opinion. However, some areas lack sufficient domain knowledge and expertise to determine which variables are best. There are also some cases where the predictive model selection method used, such as Lasso and decision tree [24]. The topic of model selection designed to make subsequent decisions has received
little attention. Researchers believe that the three types (Naive-Bayes, Decision Tree and Support Vector Machine) of methods discussed in this study show good potential to make a business decision to ensure the system running smoothly.

![Figure 5. Application of Machine Learning for Decision Making](image)

Besides that, the idea to develop Machine Learning applications in the Oil and Gas Industry is to optimise and accelerate various business processes to meet their digital transformation goals in the short and long term. Furthermore, in figure 5 shown, machine learning supports the owner to facilitates the process and accelerate the decision using that technology.

5. Conclusion
The integration of machine learning and big data will positively impact the future of the data-driven in the Oil and Gas industry. In overall, the outcome of this study to provide the current practice and future directions of research in decision making for processing big data using a model selection of machine learning methods. Research scientists, data scientists, analysts and big data practitioners must work together to create decision making. It can get more accurate and efficient data using machine learning standards and exploring more innovative in the future.

References
[1] Amir Sani. 2015. Machine Learning for Decision Making. Universite de Lille 1, English
[2] Chid Apté. 2010. The Role of Machine Learning in Business Optimization. Proceedings of the 27th International Conference on Machine Learning (ICML-10)
[3] Press, G. 2014. 6 predictions for the $125 billion big data analytics market in 2015
[4] Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.) 2013. Machine learning: An artificial intelligence approach. Springer Science & Business Media
[5] Cui, Q., Bai, F. S., Gao, B., & Liu, T. Y. 2015. Global Optimisation for Advertisement Selection in Sponsored Search. Journal of Computer Science and Technology. 30(2) 295-310.
[6] Kononenko, I. 2001. Machine Learning for Medical Diagnosis: History, State of the Art and Perspective. Artificial Intelligence in medicine. 23(1) 89-109.
[7] M. Rouse. 2011. Machine Learning Definition, http://whatis.techtarget.com/definition/machine-learning
[8] Annina Simon et al. 2015. An Overview of Machine Learning and its Applications. International Journal of Electrical Sciences & Engineering (IJESE). 1(1) 22-24.
[9] TM Mitchell, 2006. The discipline of machine learning, Carnegie Mellon University, School of Computer Science, Machine Learning Department
[10] Samuel, A. L. 1959. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development. 3(3): 211–229.
[11] Apte, C., Morgenstern, L., and Hong, S.J., 2000. AI at IBM Research, IEEE Intelligent Systems, 15(6) 51-57.
[12] Tan, P., Steinbach, M. and Kumar and V. Cluster Analysis. 2005. Basic Concepts and Algorithms. In: Introduction to Data Mining, Addison-Wesley, Boston.

[13] Azevedo, A., and Santos, M. F. 2009. Business Intelligence: State of the Art, Trends, and Open Issues. Knowledge Management and Information Sharing. 296-300

[14] Gupta S., Nikolaou M., Sapatelli L and Bravo C. 2016. ESP Health Monitoring KPI: A Real-Time Predictive Analytics Application. SPE-181009-MS. https://doi.org/10.2118/181009-MS

[15] Abdelkadir B. and Luc Q. 2014. How to use Big Data Technologies to Optimise Operations in Upstream Petroleum Industry. Presented at the 21st world petroleum congress in Moscow. Russia DOI: 10.5585/iji.v1i1.4

[16] Bravo C., Jose R., Sapatelli L. and Echevarria F. R. 2014. Applying Analytics to production Workflows: Transforming Integrated Operations into Intelligent Operations. SPE-167823-MS. https://doi.org/10.2118/167823-MS

[17] Claesens and N. L. Hjort. 2003. The Focused Information Criterion. Journal of the American Statistical Association, 98(464) 900–916.

[18] Hastie, R. Tibshirani, and J. Friedman. 2009. The Elements of Statistical Learning. Springer, New York

[19] Bousquet, S. Boucheron, and G. Lugosi. 2004. Introduction to statistical learning theory. In O. Bousquet, U. von Luxburg, and G. Rätsch, editors, Advanced Lectures on Machine Learning, volume 3176 of Lecture Notes in Computer Science, p 169–207. Springer Berlin Heidelberg

[20] Guyon, A. Saffari, G. Dror, and G. Cawley. 2010. Model selection: beyond the Bayesian/frequentist divide. Journal of Machine Learning Research. 11 61–87.

[21] Tiago M. Fragoso and Francisco Louzada Neto. 2015. Bayesian model averaging: A Systematic Review and Conceptual Classification. Departamento de Matemática Aplicada e Estatística - Instituto de Ciências Matemáticas e de Computação, c/o, Universidade de São Paulo.

[22] Wasserman. 2000. Bayesian Model Selection and Model Averaging. Journal of Mathematical Psychology. 44(1) 92–107.

[23] Bastani and M. Bayati. 2016. Online Decision-Making with High-Dimensional Covariates. Working Paper

[24] Tibshirani. 1996. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B( methodological). p 267–288. 58(1)

[25] Vahn, N. El Karoui, and A. E. B. Lim. 2014. Performance-based Regularisation in Mean-CVaR Portfolio Optimisation. Working Paper

[26] Aggarwal M. 2019. Preferences-Based Learning of the Multinomial Logit Model. Knowledge Information System. p 523–538. 59. https://doi.org/10.1007/s10115-018-1215-9

[27] Omar Besbes and Assaf Zeevi. 2012. On the (Surprising) Sufficiency of Linear Models for Dynamic Pricing with Demand Learning. Management Science. 61(4) 723-739.

[28] Cooper, William L., Tito Homem-de-Mello, and Anton J. Kleywegt. 2015. Learning and Pricing with Models that do not Explicitly Incorporate Competition. Operations research. 63 (1) 86-103.

[29] Mounica Vennapusa, Srikanth Bhuyrapuneni. 2019. A Comprehensive Study of Machine Learning Mechanisms on Big data. International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878. 7(6S2).

[30] J. Peppard and J. Ward. 2016. The Strategic Management of Information Systems: Building a Digital Strategy. Wiley. 189

[31] Sivarajah, U., Kamal, M., Irani, Z. and Weerakkody V. 2017. Critical analysis of Big Data challenges and analytical methods. Journal of Business Research. 70 263-286.

[32] Juan D. Pineda-Jaramillo. 2019. A Review of Machine Learning (ML) Algorithms used for Modeling Travel Mode Choice. Revista DYN. 86(211) 32-41.

[33] Bruckner, M., Kanzow, C., & Scheiffer, T. 2012. Static Prediction Games for Adversarial Learning Problems. Journal of Machine Learning Research. 13 2617–2654.

[34] Mert Bal, M. Fatih Amasyali, Hayri Sever, Guven Kose, and Ayse Demirhan. 2014. Performance Evaluation of the Machine Learning Algorithms Used in Inference Mechanism of a Medical
Decision Support System. Hindawi Publishing Corporation. The Scientific World Journal, Article ID 137896, 15 pages.

[35] Refaelzadeh, Payam, Lei Tang, and Huan Liu. 2007. *On Comparison of Feature Selection Algorithms*. In Proceedings of AAAI Workshop on Evaluation Methods for Machine Learning II, 34–39.