JCPT

Predicting Waterflood Responses With Decision Trees

L. FEDENCZUK, K. HOFFMANN
Gambit Consulting Ltd.
T. FEDENCZUK
University of Hawaii

Abstract

Adding new wells and new production in existing fields under Enhanced Oil Recovery (EOR) is particularly important in mature fields that are characterized by a long history of field activity. Different drilling programs, a variety of field treatments, well conversions and new injectors add many layers of complexity and uncertainty to the existing effects of geological, completion and production factors.

Surveillance and prediction of responses caused by injected fluids in fields with dozens of patterns and hundreds of wells calls for computer-based systems that estimate responses based on numerical and statistical solutions. This is especially important when geological understanding is very weak (i.e. no core, no log data).

This paper shows how results from EOR surveillance programs can be integrated with geological data. Furthermore, this paper shows how to build predictive models for production estimates based on injection responses and geology. These models support a two to three times more accurate selection of wells with high oil production during EOR than historically implemented selections.

Included in the paper are practical tips on how to select the best model and derive solutions with decision trees that are equivalent to sets of English-based rules. Solutions from decision trees are compared with solutions from logistic regression and neural networks. This comparison deals with the statistical accuracy of model predictions, interpretation ability and assisting in applying these models to support field decisions.

Introduction

The main goal of the study was to develop and test a model that would predict production performance during waterfloods. In predictive modelling, regression is traditionally applied to calculate continuous target variables[1]. Models that predict binary response variables use logistic regression[2]. A binary target variable is characterized by two events. They can be of a numerical nature (0 and 1) where zero represents a non-event and one represents an event. Alternatively, a character string with two outcomes (e.g. No and Yes) is often applied.

In the case of a continuous target variable, we may predict the fluid rates of oil, water, gas or the total fluids. The binary 0 or 1 target variable can represent a low or high production output, respectively. A production cutoff can be based on economical or engineering criteria applied to the actual rates or volumes in a specific time period.

This paper presents the principles of a numerical model development for predicting well performance during waterflood in the Pekisko B Field. The production performance predictions were based on geological and injection response parameters. The injection response parameter definition and non-numerical integration with geological data was presented in an earlier paper[3].

In this paper, the performance predictions were done for a binary target variable that identified wells with good performance. The performance was based on the actual normalized volume of production. A well's normalized production was defined as a ratio of the well production in a specific time period, to the total field production in the same time period. A binary target variable (indicator) of two levels (0 and 1) was assigned based on the normalized production. If a well production placed it in the best quartile, then the binary target value was assigned a value of one; otherwise it was assigned a value of zero.

The above target variable was predicted based on the geological and the injection response data sets. The injection response variables were identified in the earlier part of the study[3]. They included oil, water, gas and total fluid responses to the injection changes. These responses were calculated as a Spearman non-parametric correlation between the injected rates and the specific production rates (oil, water, gas and the total fluid). The geological set contained Pekisko B top subsea, Pekisko B subsea of oil-water contact and Pekisko B netpay for all wells in the field.

For this work, we have built and analyzed logistic regression, neural network and decision tree model types. These models were developed to predict the probability of high oil production. A modelling input data set contained 480 observations that were split into a learning set with 40% of the observations, a validation set with 30% of the observations and a testing set with the remaining 30% of the observations[4, 5].

This study identified large performance differences between the prediction powers of models developed with different modelling methods. Specifically, the final decision tree model outperformed the logistic regression and the neural network models. The strength of the decision tree model originates from the fact that each sequential node split (decision branch) does not have to have continuity along the boundaries between different regions or segments defined by predictive variables.

Integration of geological and response variables in a model allowed the development of rules that would support predicting a well's performance during EOR. Interpretability requirements favoured logistic regression and especially decision tree models because of their English-based nature of rules. Models based on neural networks did not prove superior to other model types and, at best, they provided limited interpretation support of predictions.

PEER REVIEWED PAPER ("REVIEW AND PUBLICATION PROCESS" CAN BE FOUND ON OUR WEB SITE)

Injection Responses

Our injection response evaluation is based on the injected and produced rates. The methodology was developed from experience with field studies for Golden Lake, Swan Hills, Midale, Valhalla, Goose River, Cactus Lake, Mirage and Pekisko B[3]. Related discussions can also be found in earlier publications[6, 7]. The technique is applicable to vertical and/or horizontal wells for injection surveillance and optimization. It can play an essential role in studies of underperforming fields or acquisition targets. Furthermore, the same technique can detect communication between producers and can help in designing new waterfloods.

Relationships between produced rates of oil, water, gas or the total fluid, and the injected rates of water can indicate fluid communication through a reservoir. However, typical oil fields can exhibit complex geology across a field or across patterns, accidental schedules of wells and/or random changes in the injection and production rates. Together with the shear volume of data, manual analyses may lead to ambiguous and biased associations between producers and injectors. Our methodology technique provides a rigorous and unbiased approach. It is based on the Spearman rank correlations between the injected and produced rates, over a period of time series[8]. These correlations, and the time lags between the injection and the associated production rates, allow us to compress these series of rates into a set of simple parameters. We use these parameters to estimate oil, water, gas, and total fluid responses for every combination of injector and producer.

In regular patterns with vertical wells, the correlations (oil, gas, water and total fluid responses) and associated time lags can be presented in the form of a single or composite star and spider diagrams. In an integration process, sets of composite diagrams were overlaid with contour maps of formation tops and netpay[3]. These presentations helped find the significant relationships between the producers' responses and the underlying geology and helped in understanding field behaviour. These composite diagrams can also help to:

- evaluate sweep efficiency;
- select areas for infill programs;
- identify ineffective injectors;
- identify producers without support;
- better estimate the production allocation;
- find areas with fluid losses; and,
- develop communication/correction maps for reservoir simulation.

For a detailed description of the methodology and visualization techniques, please refer to earlier papers[3, 6, 7].

Presented here is the numerical integration of the waterflood responses, geology and other available data sources. This gives a more formalized approach than previously applied by overlaying contour maps with different parameters[3, 9]. Furthermore, it enforces consistency, removes the analyst's bias and allows for generating English-based rules, which can be applied in field decisions. The advantage of the approach increases for fields with a large number of injectors and many years of history.

Model Development

The modelling process described in this section is characteristic of a data-driven model development[9, 10]. Custom programs were used to load, format, summarize and transform data from the external data sources. The final data set(s) formed a project database. Flat files, Excel tables, archived files and databases provided the required interface for production and geological data sources.

Initially, a set of programs was developed which called on a variety of statistical procedures. These procedures included regression, logistic regression and discriminant analysis[1, 4]. This project involved numerous iterations, which included data manipulations, data analysis, model development and reviews of their performance.

We developed and refined three different model types, which were based on regression, neural networks[11, 12] and decision trees[13, 14]. The selection of variables, diagnostics and interpretation were heavily used to justify each of the development steps and directed further research. Descriptive statistics played a significant role in summarizing data, generating categorical variables and defining normalized parameters. These additional categorical variables had a triple purpose. First, they represented initial segmentation based on 'known' geological knowledge and observed distributions. Secondly, some of these categorical variables were generated for the model performance testing, verification of performance segments based on residuals and verification of untested hypothesis. Thirdly, some of the categorical variables represented a hierarchy of dimensions (geography, well type and time) and were designed to support multidimensional reporting of historical data and model predictions.

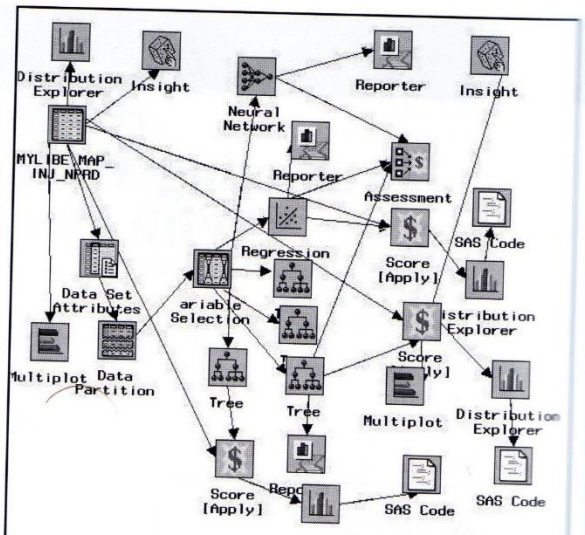


FIGURE 1: Project flow diagram.

Project Diagram Flow

A project flow diagram for the well performance prediction or ranking (based on normalized oil production) is presented in Figure 1. At one point, this diagram contained three decision tree models and the same number of other model types. These models were used to compare one-type models. In particular, they compared different neural network or decision tree models. For example, we tested a variety of neural network models with a different number of hidden layers and direct links between the input and output layers. Similarly, the decision tree depth, the splitting criteria, splitting variables, the business interpretability and the misclassification rates were compared before the best tree model was selected.

The process flow started with the Input Data Source node. This node mapped data and assigned additional variable attributes that were required by modelling nodes (Regression, Neural Network and Tree). These attributes included the model role for each variable (i.e. target, input, rejected) and were changed from their default assignments when required. Tables of statistics and histograms were reviewed for interval and class variables.

The second node was the Attribute node, and it marked variables to be used. Next, this node assigned a role, a type (character or numerical) and additional attributes for each variable. A variable's role could be an ID, target, input or rejected label, while the

TABLE 1: Profit matrix.

	Predicted Outcome	
Actual Outcome	**1 (Good)**	**0 (Poor)**
1	5,000,000	0
0	0	0

TABLE 2: Constant cost matrix.

Decision	Cost
1	500,000
0	0

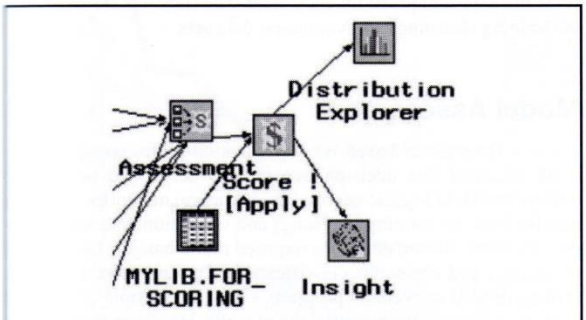


FIGURE 2: The Assessment and Scoring nodes.

variable's measurement could be assigned to unary, binary, nominal, ordinal and interval.

Furthermore, models in this study were optimized to maximize profit based on a constant cost and expected profits associated with each decision. This required defining a 'Profit Matrix' and a 'Constant Cost Matrix.' The first was a 2 × 2 matrix that represented the expected profit (see Table 1) based on actual and predicted outcomes (1 = good well, 0 = poor well). The Constant Cost Matrix in Table 2 contained two rows with costs based on two decisions (1 or 0).

Next, the Partition node performed data sampling into learning, validation and testing sets. These data sets resulted from a combination of the user-defined sampling and the random sampling. Three subsets, Train, Validation and Test, were selected from the original data set of 480 observations.

The Variable Selection node assisted in reducing the number of inputs by setting a rejected status of all input variables that were not related to the target. In some cases, this automatic selection was overridden by assigning the input status to a rejected variable or the rejected status to an input variable. The subset of the most important inputs was then evaluated in more detail by one of the modelling nodes.

The next vertical layer of nodes in Figure 1 consisted of the Regression (logistic for the binary target) node, the Tree node and the Neural Network node. These modelling nodes performed all of the steps required to find the most optimal model for the specific model type.

Finally, the Score node (see Figure 1) was used to generate predictions from a trained model and a new input data set. This node applied each model's formula to the 'unknown' data set. The predictions were accompanied by assessment statistics. Each of the modelling nodes in Figure 1 was connected to its own Reporter node.

The Assessment node (Figure 1 and Figure 2) compared models and prediction diagnostics for all modelling nodes. A more advanced comparison could be facilitated with a set of advanced charts for lift, profit, return on investment (ROI), receiver operating curve (ROC) and response threshold chart[5, 15, 16]. A direct link between a specific modelling node and the Assessment node was applied to reassure the user's model selection (Figure 2).

At different steps, two more nodes were applied to review the data and results. First, the Distribution Explorer node enabled visual exploration of large volumes of data. The node was used primarily in the exploration phase to uncover patterns and trends and to reveal extreme values in the database. Next, the Insight node allowed exploring and analyzing the data through graphs and analyses that were linked across multiple windows.

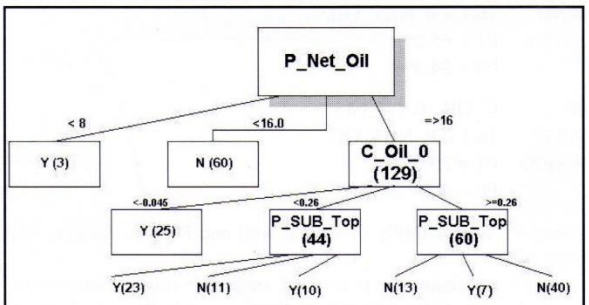


FIGURE 3: Tree diagram; subdivision levels = 3; max number of splits = 3.

Decision Tree Modelling

Decision trees are well suited for clustering and classification tasks. Decision trees classify data by applying a series of simple rules. Each rule assigns an observation to a class based on one specific parameter in a recursive fashion. The resulting classes are individually divided into new classes, based on new splitting parameters and rules applied to these parameters. All subdivided and non-subdivided classes are called nodes, and create the hierarchical structure of the decision tree. This rule-based recursive splitting process generates branches that can vary in depth. The depth, in turn, corresponds to the number of subdivision levels. The original class contains the entire data set and is called the root node of the tree. The final nodes that are not subdivided are called the leaves. Such hierarchical structure corresponds to an inverted tree where the root is on the top and the leaves are at the bottom of the tree structure.

When being developed with the training set, trees divide a population into segments with similar characteristics. In our case, we wanted to find out which, of a long list of attributes (geological and response parameters), were the best predictors of a well's performance, what rules they followed and where in the tree we should apply them. In general, a decision tree applies the same decision to each observation that trickles through the set of rules from the tree root and ends up in the same leaf node. This means the same classification (good/poor) or the same production value (e.g. 8.31 m^3/day) is associated with all observations in the same leaf node.

Figure 3 presents a tree example with the corresponding values of parameters that were used to split the nodes at different levels of the tree subdivision. At first, the algorithm might have determined that the attribute with the most impact was P_Net_Oil (Producer's Pekisko B net oil), and then might have decided to split the population into three groups or clusters based on the net pay <8, <16 and ≥16. The next most important splits, in order, might have been C_Oil_0 (zero lag normalized oil response) and P_SUB_Top (Producer's Pekisko B top subsea). Symbols 'Y' or 'N' in Figure 3 identify good and poor classification clusters. These leaf nodes represented the nodes that were not subdivided. Numbers in brackets show the number of observations in each leaf node (non-divided bin). A detailed description of non-lagged and lagged waterflood responses was presented in earlier papers[3, 6, 7].

A final decision tree model (Figure 3) can be used for classifying a new well or newly converted or treated wells from the geological parameters and the instantaneous oil response (C_Oil_0 – response at time lag = 0). This model assigns wells to two risk groups of good and poor producers (Y or N). An example of a pseudo code that corresponds to some portions of this decision tree is presented below:

```
IF      16.1 ≤ P_NET_OIL < 17.9
AND     –1250.35 ≤ P_SUB_TOP < –1,244.95
AND     0.26 ≤ C_OIL_0
THEN    P1 = 100.0%
        P0 = 0.0%

IF      P_SUB_TOP < –1,252.21
AND     –0.045 ≤ C_OIL_0 < 0.26
```

```
AND     16.1 ≤ P_NET_OIL
THEN    P1 = 65.2%
        P0 = 34.8%

IF      C_OIL_0 < -0.045
AND     16.1 ≤ P_NET_OIL
THEN    P1 = 20.0%
        P0 = 80.0%
```

where P1 is probability of a good well and P0 is probability of a poor well.

When a decision tree is verified, such code can be easily implemented in any software package that is used in the petroleum industry. In this specific example, the estimated classification will be based on the posterior probability of a good well (P1) and a poor well (P0). With a 50% threshold cutoff value, a user's decision will be estimated from a simple formula:

IF (P1 ≥ 50%) THEN Good ELSE Poor.

There are two opposing activities during the tree model development. First, an algorithm generates a full-grown tree by a recursive node splitting, and the second prunes explicit nodes or sub-trees in order to retain the most optimal tree[17, 18]. A recursive splitting of nodes during a tree construction is based on the strength (statistics) of the splitting rules:

- If the Chi-square or the F test criterion is selected, then the computed statistic is the LOGWORTH = -log (p-value from Chi-square or F test).
- If the Entropy or Gini reduction criterion is selected, then the computed statistic is the WORTH, which measures the reduction in variance for the split[5].

Larger values for both LOGWORTH and WORTH are better. The method is recursive because each set of new nodes results from the splitting of a previously divided node. After a node is split, the newly created nodes are considered for splitting. This recursive process ends when no node can be split any further.

Table 3 and Table 4 show two different geological and waterflood response criteria used to evaluate competing node splits. Such tables were used during the interactive development of the decision trees.

TABLE 3: Competing splits for a tree with three branches. Splitting criterion based on Gini test.

Variable	Logworth	Groups
GMC_LAG	2.654	2
OMC_LAG	2.474	3
RESPONSE	2.447	2
C_GAS	2.35	2
P_NET_OIL	1.755	3

TABLE 4: Competing splits for a tree with two or three branches. Splitting criterion based on Chi-square or F.

Variable	Worth	Groups
P_NET_OIL	0.160	3
OMC_LAG	0.149	3
C_GAS	0.138	3
P_SUB_TOP	0.129	3
GMC_LAG	0.116	3

TABLE 5: Example of a tree node statistics.

Target Values	Training Data	Validation Data	
1	86.7%	80.0%	% for each
0	13.3%	20.0%	target level
1	13	4	Count for each
0	2	1	target level
Total	15	5	
Decision	1	.	
1	3,833,333	4,000,000	
0	0.13333	0.2	Expected profit

In addition, different sub-tree methods determine which sub-tree is selected from the fully-grown tree. This process can be based on whether or not the profit/loss matrix is used for a split search. Table 5 shows an example of tree node diagnostics for each target level in per cent, the corresponding count, the total count and the overall decision level associated with a specific node. The last two rows show the expected profit for each level. The statistics are shown for the training (learning) and validation data sets.

Model Assessment

Early diagnostics based on classification tables (confusion tables) indicated that decision tree models performed better than models based on logistic regression and neural networks. This was true for both the training (learning) and validation data sets. However, business-driven decisions required more than just two rates of the correct and erroneous classifications. In a non-discriminatory drilling or well conversion program, single or multiple criteria can be used to identify the potential list of wells. However, the program cost can be lowered substantially if we identify a much smaller portion of wells that are most likely to respond to the implemented waterflood with the right response type and higher oil rates.

A well's classification in this project was based on a pre-selected cutoff applied to the estimated posterior probability. The following pseudo-code shows this logic:

```
IF (posterior probability ≥ cutoff)
THEN Good Well
ELSE Poor Well
```

More advanced analysis and identification of the probability percentage cutoff was facilitated with lift curves. In a lift chart (also known as a gains chart) for a non-binary target, all observations from the scored data set are sorted from the highest to the lowest production probability. For a binary target, the scored data set is sorted by the posterior probabilities of the event level (production in the highest quartile) in descending order. Then the observations are grouped into deciles.

Figure 4 shows an example of a cumulative percent response lift chart for three models. In this chart, the target production index is sorted from left to right, starting with wells that are most likely to produce. This likeliness was estimated based on the posterior probability of the target event level equal to one (high production) as predicted by each model. The sorted group is lumped into 10 percentiles along the X-axis. The left-most percentile is the 10% of the

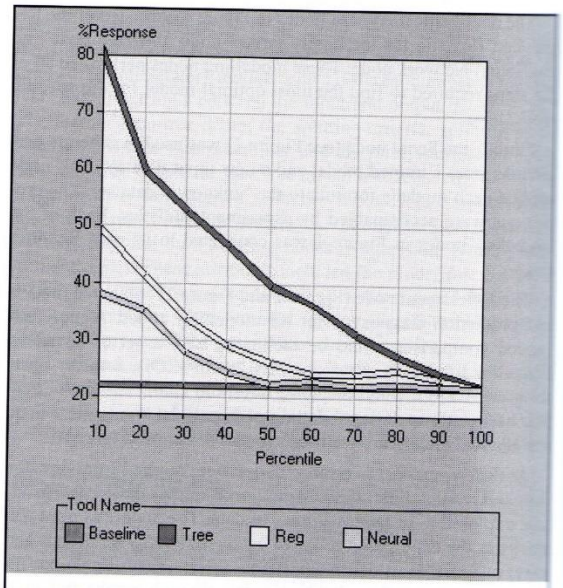


FIGURE 4: Cumulative response curves.

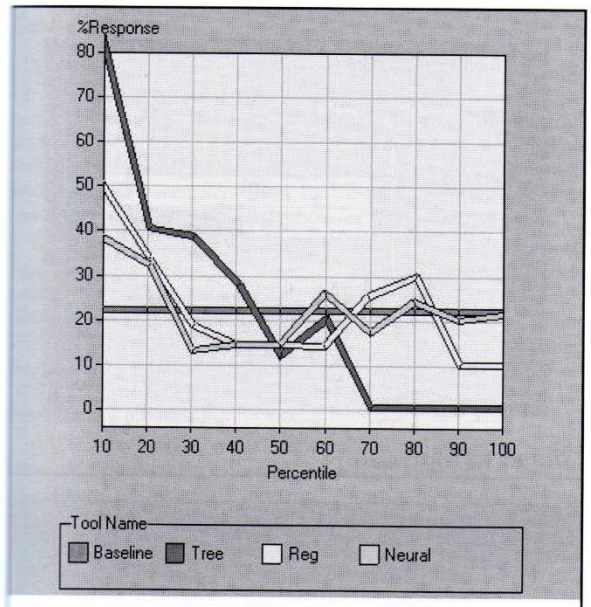


FIGURE 5: Non-cumulative response curves.

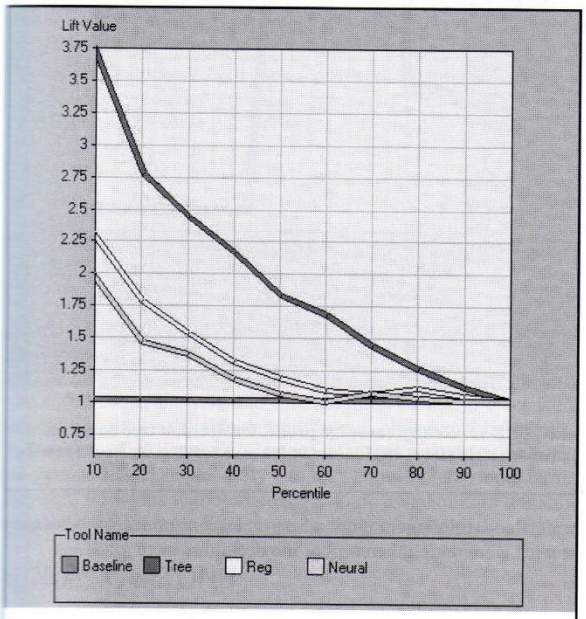


FIGURE 6: Cumulative lift curves.

wells that are the most likely to produce (highest production probability). The vertical axis represents the predicted overall cumulative percentile of good producers in the selected deciles along the X-axis. Thus, if we drill/convert all wells (100%), the response (percentage of good wells) will be equal to the success observed in the whole sample (22 – 23%). However, if we go after the best 10%, 20% or 30% of all wells then the success rate will be around 83%, 70% and 52%, respectively. Figure 5 presents non-cumulative response curves, which show the percentage of good producers in each decile. A baseline in this figure shows an average percentage of wells with good performance in the original sample.

The next two figures show the lift curves in a relative scale where the baseline corresponds to one (historical success rate). Figure 6 shows the cumulative lift curves, which correspond to the three models (tree, logistic regression and neural net). A lift curve shows the model's effectiveness relative to a baseline, which shows an overall (average) historical success rate (horizontal line).

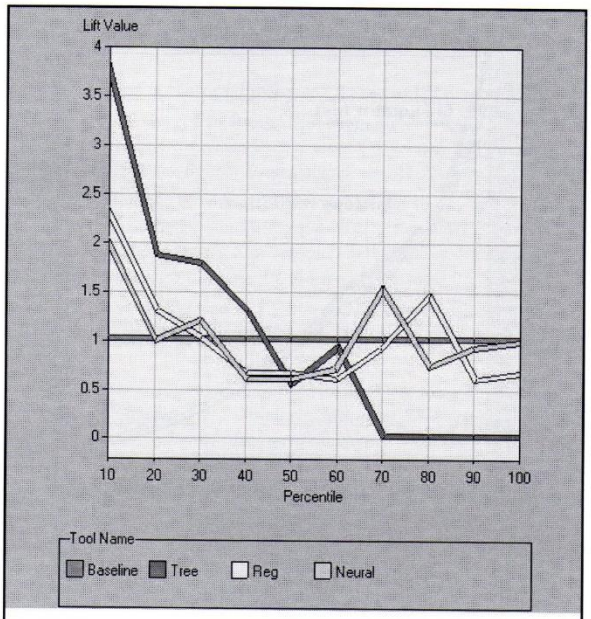


FIGURE 7: Non-cumulative lift curves.

Non-cumulative lift curves (shown in Figure 7) enhance the visual comparison of the model's performance in each decile.

The non-cumulative lift curve for the decision tree in Figure 7 shows that beyond the fourth decile, most of the best producers would be selected and the rest of the wells should perform well below the overall average. The non-cumulative lift curve for the tree model in Figure 7 shows nearly a two to four times better success rate than historically observed in the field. This range of improvement in the well selection would be achieved if the tree model was implemented and used to select only 20% and 10% of the best wells, respectively.

Both sets of lift curves showed that the logistic and neural models significantly underperformed relative to the decision tree model. The non-cumulative lift curves in Figure 7 showed that the best model performance dropped fast from around 4 to 1.8 between the first and the third best deciles.

Different node splitting criteria can make a difference in most instances. Figure 8 shows a comparison of lift curves between the three decision trees with different node splitting criteria. In this specific case, applying the Chi-square test to evaluate the node splitting criterion provided the best lift in the first two deciles[5].

Many modelling decisions, as well as the model selection, depended on the misclassification rates. Figure 9 shows a confusion (classification) chart with agreements between the actual and predicted counts for the tree model at the 50% threshold value. This diagram helped with verifying the agreement between the actual and the predicted classes at different threshold levels. The threshold level is the cutoff that was applied in classifying observations based on the evaluated posterior probabilities. If a predicted score was below the threshold value, then the predicted production class was assigned to zero (production below the desired level); otherwise the class was assigned to one (good production).

Well identification efforts and drilling programs have associated costs and returns on investment for each case of four outcomes between the predicted and the actual outcomes. Figure 10 presents a profit matrix for four hypothetical outcomes. A simple (0/1 or N/Y) decision schema had two cases of misclassification and two cases of correct classification. The assigned fix profit was based on a simple principle that a successful prediction (identified as a good prospect) would generate 10 units (in millions of $), less 0.0 units of the fixed costs (see 1/1 cell with return = 10). A non-successful well pick, which was classified as a good producer, had a negative return related to the fixed cost (-0.5). The predicted non-events were classified in a similar way, where 0 was assigned for

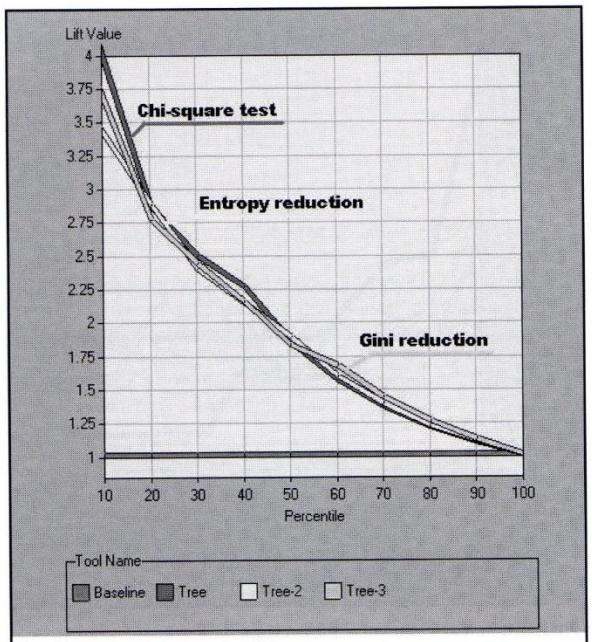


FIGURE 8: Cumulative lift curves for three decision trees with different node splitting criteria.

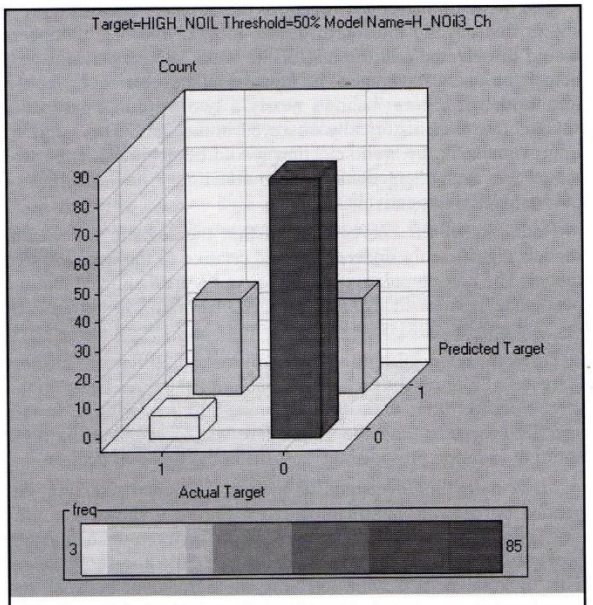


FIGURE 9: Threshold chart for a tree model at threshold of 50%.

the 1/0 case (missed revenue), and 0 for the 0/0 case (the correct prediction of the non-event). This was one of many scenarios that could be used to test the model performance, stability and sensitivity. The presented values do not reflect actual values in this particular field.

The corresponding profit (return) chart in Figure 11 shows the relationship of the estimated return versus the classification threshold value (IF posterior probability ≥ threshold, THEN class = 1). This diagram shows that thresholds in the range of 15 – 50% should generate the highest average return. It shows that the best average return and the highest total production volumes could be achieved at the 15% threshold value. This translates to selecting most of the wells (IF posterior probability ≥ 15%, THEN Good). The above example characterized a relatively successful

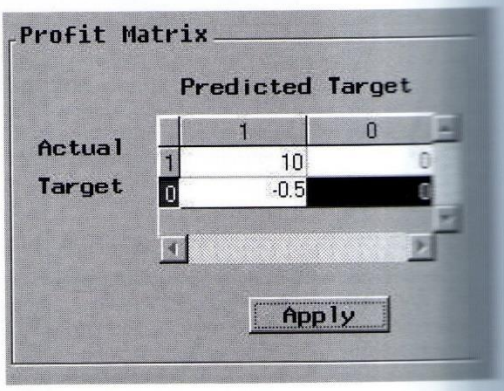


FIGURE 10: Profit matrix.

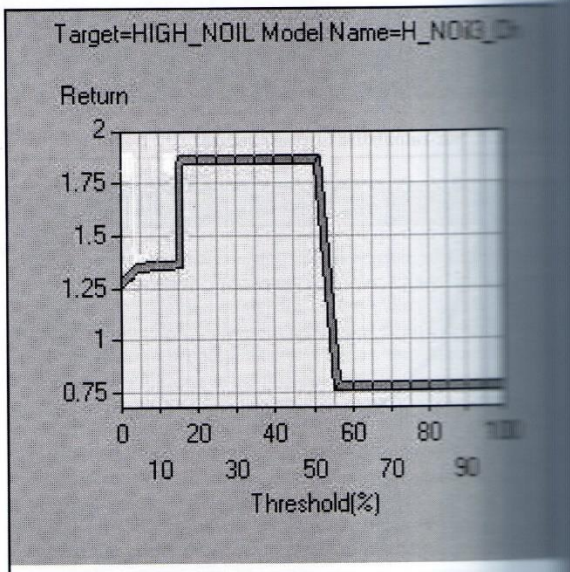


FIGURE 11: Return (average profit) for decision tree [illegible] profit matrix (1/1 = 10M; 1/0 = 0; 0/1 = -0.5; 0/0 = 0); Profit [illegible] from Figure 10.

waterflood implementation where a large amount of [illegible] studies were undertaken before the decisions were made.

Decision Tree Versus Neural Networks

Neural networks have been utilized in a variety of studies [illegible] optimism fueled by the origin of this tool and from publications [illegible] cessful applications. However, in this study, the neural networks were not able to prove their strength[19-22]. In our implementation of the neural network model (Figures 4 to 8), we added [illegible] ables, hidden layers and direct links between the input and [illegible] layer. None of these attempts had any positive effect, and the [illegible] network models produced poorer predictions. Thus, the final [illegible] utilized a formula that was based on the decision tree model.

Furthermore, neural network development requires [illegible] statistical analysis in order to understand the data and the [illegible] flows. Most practitioners apply the stepwise regression, the [illegible] ward regression and the decision tree variable selection [illegible] plying neural network modelling. Finally, neural network [illegible] cannot be directly applied in business interpretation [illegible] which in some cases can eliminate the neural model from [illegible] eration. Therefore, only significantly better performance [illegible] tion rates could justify the neural network model implementation.

Conclusions

The case described in this paper is not complete. Specifically, integrated the waterflood responses with a few geological parameters. Additional data sources (e.g. completion, seismic, petrophysical) should be integrated with the waterflood responses. This study shows that numerical integration of geological and waterflood response parameters allow for the prediction of oil production from enhanced recovery processes.

Different model types were built, which included decision trees, regression models and neural networks. These predictive models were developed for a binary target variable that indicated a well's performance at two levels (0/1 or Poor/Good). The indicator variable was derived from the normalized oil production and identified the best 25% of all wells in the whole field. The normalized oil production was characterized by a well's production relative to the whole field's production.

Logistic regression, neural network and decision tree models were developed and compared. The decision tree model was selected based on the best performance. An advantage of the decision tree model over the other types of models was that it could produce rules that represented interpretable English-rules or logic statements. For example, "If netpay is greater than 5 m and the lag zero response is negative, then oil production will be in the top 25% of best production with a probability of 80%."

Statistical diagnostics based on the model verification process showed that selecting wells based on models that use geological and fluid communication parameters resulted in a success rate of two to four times better than by traditional methods.

Furthermore, we showed how a profit matrix might be used to combine model prediction with the impact (cost) of all classification outcomes (true positive, false positive, true negative and false negative predictions).

Model-based decision tools require data collection, data mining, appropriate model selection, model diagnostics and information visualization. Furthermore, simplicity, interpretability, maintenance requirements and the stability of models should influence the model development approach.

Disclaimer

All data sets, analysis and interpretation in this project were based on publicly available data and results already published. The decision tree and its scoring logic that are presented in this paper are provided as examples only and they should not be used in actual field decisions without extensive knowledge of data transformations, inclusions, exclusions and time period analysis.

NOMENCLATURE

OIL_0	=	oil response at time lag = 0 (instantaneous response)
P	=	probability of poor performance (1-P)
G	=	probability of good performance (above a threshold)
NET_OIL	=	Pekisko B net oil
PEK_TOP	=	Pekisko B top subsea

REFERENCES

DAVIS, J.C., Statistics and Data Analysis in Geology; *2nd Edition, John Wiley & Sons, Inc., Hoboken, NJ, 1986.*

SAS INSTITUTE INC., Logistic Regression Examples Using the SAS System; *Version 6, First Edition, SAS Institute Inc., Cary, NC, 1995.*

FEDENCZUK, L., PEDERSEN, P. and MARSHALL, M., Analyzing Waterflood Responses for Pekisko B; *Journal of Canadian Petroleum Technology, Vol. 40, No. 6, pp. 29-35, June 2001.*

SAS INSTITUTE INC., SAS/STAT User's Guide; *Version 6, Fourth Edition, SAS Institute Inc., Cary, NC, 1989.*

SAS INSTITUTE INC., Data Mining Using Enterprise Miner Software: A Case Study Approach; *First Edition, SAS Institute Inc., Cary, NC, 2000.*

FEDENCZUK, L., PEDERSEN, P. and MARSHALL, M., Analyzing Waterflood Responses for Pekisko B.; *paper No. 99-46 presented at the CSPG and Petroleum Society Joint Convention, Digging Deeper, Finding a Better Bottom Line, Calgary, AB, 14-18 June 1999.*

FEDENCZUK, L. and HOFFMANN, K., Surveying and Analyzing Injection Responses for Patterns With Horizontal Wells, *paper SPE 50430 presented at the SPE International Conference on Horizontal Well Technology, Calgary, AB, 1-4 November 1998.*

BERENSON, M.L. and LEVINE, D.M., Basic Business Statistics: Concepts and Applications; *Prentice Hall, Inc., Englewood Cliffs, NJ, 1983.*

FEDENCZUK, L. and HOFFMANN, K., Data Integration and Analysis for Optimal Field Development; *paper 97-45 presented at the 48th Annual Technical Meeting of the Petroleum Society of CIM, Calgary, AB, 8-11 June 1997.*

BERRY, M.J.A. and LINOFF, G., Data Mining Techniques: For Marketing, Sales and Customer Support; *John Wiley & Sons, Inc., Hoboken, NJ, 1997.*

BISHOP, C.M., Neural Networks for Pattern Recognition; *Oxford University Press, New York, NY, 1995.*

BIGUS, J.P., Data Mining with Neural Networks: Solving Business Problems—From Application Development to Decision Support; *McGraw-Hill, Inc., Hightstown, NY,* 1996.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J., Classification and Regression Trees; *The Wadsworth Statistics/ Probability Series, CRC Press LLC, Boca Raton, FL, 1984.*

APTE, C. and WEISS, S.M., Data Mining With Decision Trees and Decision Rules; *Future Generation Computer Systems, Vol. 13, No. 2, pp. 197-210, November 1997.*

METZ, C.E., Basic Principles of ROC Analysis; *Seminars in Nuclear Medicine, Vol. 8, No. 4, pp. 283-298, October 1978.*

HANLEY, J.A. and MCNEIL, B.J., The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve; *Radiology, Vol. 143, No. 11, pp. 29-36, April 1982.*

FEELDERS, A., Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation?; *Lecture Notes in Computer Science, Vol. 1704, pp. 329-334, 1999.*

MEHTA, M., RISSANEN, J. and AGRAWAL, R., MDL-Based Decision Tree Pruning; *Proceedings of the First International Conference on Knowledge Discovery in Databases and Data Mining (KDD-95), Montreal, PQ, 20-21 August 1995.*

ZHANG, L. and ZHANG, B., Neural Network Based Classifiers for a Vast Amount of Data; *Lecture Notes in Computer Science, Vol. 1574, pp. 238-246, 1999.*

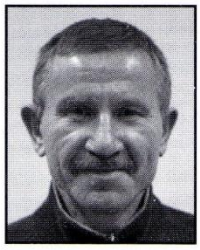
MONZURUR RAHMAN, S.M.. YU, X. and MARTIN, G., Neural Network Approach for Data Mining. Progress in Connectionsist-Based Information Systems; *Proceedings of the 1997 International Conference on Neural Information Processing and Intelligent Information Systems, Vol. 2, pp. 851-854, Springer, 1997.*

DENG, C. and XIONG, F., Neural Method for Detection of Complex Patterns in Databases; *Lecture Notes in Computer Science, Vol. 1574, pp. 258-262, 1999.*

JACOBSEN, C., ZSCHERPEL, U. and PERNER, P., A Comparison between Neural Networks and Decision Trees; *Lecture Notes in Computer Science, Vol. 1715, pp. 144-158, September 1999.*

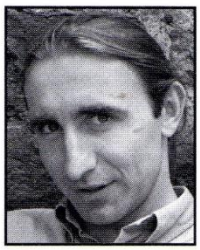
Provenance—Original Petroleum Society manuscript, <**Title**> (2002-028), first presented at the 3rd Canadian International Petroleum Conference (the 53rd Annual Technical Meeting of the Petroleum Society), June 11-13, 2002, in Calgary, Alberta. Abstract submitted for review December 3, 2001; editorial comments sent to the author(s) December 11, 2006; revised manuscript received January 18, 2007; paper approved for pre-press January 18, 2007; final approval April 11, 2007.

Authors' Biographies



Leon Fedenczuk is a Senior Consultant with Gambit Consulting Ltd. He obtained his B.Sc. and Ph.D. degrees from the University of Calgary. His expertise lies in stochastic and deterministic modelling methods, data mining, computer applications, dimension reduction and up-scaling techniques in the petroleum industry. He has been developing new ways to improve decision support in the petroleum industry since 1982. Before joining Gambit Consulting, his career included nine years with Canadian Hunter Exploration.

Kristina Hoffmann is President of Gambit Consulting Ltd. She holds a B.Sc. in physics and an M.Sc. in physical chemistry. Her career includes eight years with Canadian Hunter as a Core Analysis Coordinator and 16 years as a consultant for an array of petroleum companies working on assessments of oil and gas fields in Canada, the US, South America and Africa. She is a recognized expert in reservoir characterization based on core studies, evaluation of formation damage and field performance during completions, waterfloods and stimulation.



Tom Fedenczuk is a Ph.D. graduate student in geology and geophysics at the School of Ocean and Earth Science and Technology/ Hawaii Institute of Geophysics and Planetology (SOEST/HIGP), University of Hawaii. He specializes in GIS, visualization, geomorphology and hydrogeology.