PETROLEUM SOCIETY

*PAPER 2006-201*


# Predicting Waterflood Responses for Pekisko B.

LEON FEDENCZUK, KRISTINA HOFFMANN, TOM FEDENCZUK

GAMBIT CONSULTING LTD.

This paper is to be presented at the Petroleum Society's 7th Canadian International Petroleum Conference (57th Annual Technical Meeting), Calgary, Alberta, Canada, June 13 – 15, 2006. Discussion of this paper is invited and may be presented at the meeting if filed in writing with the technical program chairman prior to the conclusion of the meeting. This paper and any discussion filed will be considered for publication in Petroleum Society journals. Publication rights are reserved. This is a pre-print and subject to correction.

## ABSTRACT

Adding new wells and new production in existing fields under EOR is particularly important in matured fields that are characterized by a long history of field activity. Different drilling programs, variety of field treatments, well conversions, and new injectors add many layers of complexity and uncertainty on top of 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 (no core, no log data).

This paper shows how results from EOR surveillance programs can be integrated with the geological data. Furthermore, this paper shows how to build predictive models for the production estimates based on the injection responses and geology. These models support 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 assist in applying these models to support field decisions.

Finally, the new results presented here have come from the two versions of the improved decision tree model. In the first version a better version of the model was built. This model was characterized by a lift curve above four (400% improvement in the first decile). Such a result represents a significant improvement in comparison to the model from the 2001 report and presented during CIPC 2002 conference. In the second

version new core and completion data set were added when building the model.

**Keywords**: EOR, Waterflood responses, Modeling, Decision Trees, Neural Networks, Logistic Regression, Optimizing Injection, Fluid Communication, Field Surveillance.

## INTRODUCTION

The project's main goal was to develop a predictive model for production rates during waterfloods. In predictive modeling, regression is traditionally applied to predict continuos 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 numerical nature *0* and *1* where zero represents 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 continuos target variable, we may predict the fluid rates of oil, water, gas, or the total fluids. The binary 0/1-target variable can represent a low or high production output respectively. A production cut-off can be based on economical or engineering criteria applied to the actual rates or volumes in a specific time period.

This paper presents the history of a model development for predicting waterfood responses in the Pekisko B. field[3]. These predictions were done in two forms. The first form predicted the actual normalized volume of production. The 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. In the second approach a high-normalized production binary indicator variable of two levels (*0 and1)* was defined based on a cut-off. 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 value of zero.

Both types of the target variables (continuous and categorical) were predicted based on the geological and the injection response data sets. The injection response variables were identified in the earlier part of the project[3]. They included oil, water, gas, and total fluid responses to the injection changes. These responses were calculated as 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.

In each case, we have built regression, neural network, and decision tree model types. In the case of the binary target variable, a logistic regression model was developed to predict the probability of a high oil production. A modeling 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.

This study identified large performance differences between the prediction powers of models developed with different modeling 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. We identified that special care should be taken when developing the tree models. They often become unstable when there are many variables that compete in the splitting of a specific node.

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 requirement favored regression, 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 models and they definitely did not support any interpretation of predictions.

In most cases, the models developed in this project were based on a pre-selected subset of variables that provided the best quality predictive models. In the initial stages of the model development, these subsets of variables were derived with stepwise selection regression, backward elimination regression, or decision tree techniques [4,5].

### INJECTION RESPONSES

This study is based on new developments for analyzing injection responses in patterns for waterflood optimization. Our method is based on the injection and production rates history. The principles of the methodology were recently published in the JCPT Journal (June 2001)[3]. Earlier publications included paper 99-46 during the CIM Conference (June 1999)[6] and an SPE paper during the Conference on Horizontal Well Technology (October 1998)[7]. The technique is applicable to vertical and/or horizontal wells for injection optimization. It can however play an essential role in studies to locate the under-performing fields that may represent the acquisition targets. Furthermore, the same technique can detect communication between producers and can help in designing new waterfloods. Vertical and especially horizontal wells can respond to a dozen or more of vertical injectors. Understanding fluid communication between the production well and the surrounding injectors is essential to estimating the effectiveness of the waterflood and helps predict responses to the waterflood. Armed with the understanding of responses, we can optimize injection patterns, improve production rates, and achieve more efficient oil recovery.

Comparing the produced rates of oil, water, gas and the total fluid to the injected water demonstrates fluid communication through a reservoir. However, typical oilfields 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 correlation between the injected and produced rates over a period of time series[8]. This correlation 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 estimate oil, water, gas, and total fluid responses.

This study showed, that better characterization of fluid communication with the Spearman rank correlation, can be achieved when differencing is applied to the input time series for the injection rates and rates of oil, water, gas, and the total fluid. These response parameters are estimated 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 diagram.* Furthermore, the same parameters can be presented on *composite spider graphs*, which show the responses at both the short and the long distance scale.

In an integration process, sets of composite diagrams can be overlaid with contour maps of facies or netpay maps. These presentations help find the significant relationships between the producer's responses and the underlying geology and help in understanding field behavior. They can also help to evaluate sweep efficiency, select areas for the infill programs, identify ineffective injectors, identify producers without support, better

estimate the production allocation, find areas with fluid loses, and develop communication/correction maps for reservoir simulation. For detailed description of the methodology and visualization techniques please refer to previously quoted papers[3,6,7].

The advantage of the approach is the capability to do a quick analysis and interpretation of fields under the waterflood, with a large number of injectors and many years of history. This methodology was developed from our experience in field studies for Golden Lake, Swan Hills, Midale, Valhalla, Goose River, Cactus Lake, Mirage, and Pekisko B.

## MODEL DEVELOPMENT

The modeling 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 to production and geological data sources.

Initially, a set of programs was developed, which called a variety of procedures. These procedures included regression, logistic regression, and discriminant analysis[1,4]. Many programs were modified and different options were enabled, as they were required. The log and output printouts were created and some of them were saved as files for future reference. This project involved numerous iterations, the maintenance of programs, outputs, and options. Developing a good model is no longer an isolated one time project. Usually it is only the first step in the development of a set of models. The best of them may be used to develop a production decision support system (DSS). Such systems usually are vehicles of innovation, cost reduction, and improved decision support. The best model type selection becomes more obvious when one compares solutions from different model types and finds drastic differences in their performance.

In the second part of this project, we developed and refined three different model types. These models were based on regression, neural networks[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 activity.

The initial data loads and data cleaning were not repeated. However, we performed new data summarization, normalization, and segmenting. Designing data normalization and compression often represents the most important part of any data-mining project.

Descriptive statistics played significant role in summarization processes, generation of categorical variables, and defining normalized parameters. These additional variables had 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 them represented a hierarchy of dimensions (geography, well type, and time) and were designed to support multidimensional reporting of historical data and model predictions.

## 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 two to three decision tree models and the same number of other model types. They were used to compare between a series of one-type models. In particular, they allowed for comparison of different neural network or decision tree models. For example, we tested a variety of neural network models with different number of hidden layers or/and with direct links between the input and output layers. Similarly, the decision tree depth, the splitting criteria, splitting variables, the business interpretability, and the miss-classification rates were compared before the best tree model was selected.

The process flow started with the Input Data Source node. This node mapped data from a set with the pre-summarized data and assigned additional variable attributes that were required by modeling nodes (Regression, Neural Network, and Tree). These attributes included the model role for each variable (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 the Attribute marked variables to be used. Next, this node assigned a role, a type (character or numerical), and additional attributes of each variable. A variable's role could be id, target, input, and rejected, while the variables measurement cold 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 'Profit matrix' and 'Constant Cost Matrix'. The first was a 2*2 matrix that represented the expected profit (see Table 1) based on actual and predicted outcomes (1/0 Good/Poor well). The constant cost matrix in Table 2 contained two rows with costs based on two decisions (1/0).

Table 1. Profit matrix.

| Predicted<br>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 |

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. Stratified sampling and user-defined sampling were tested in this node as well.

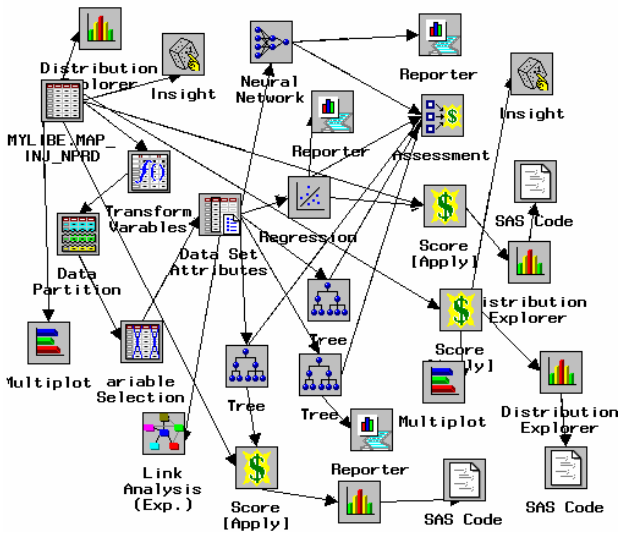

Figure 1. Project flow diagram.

Many data mining databases have hundreds of potential model inputs (independent or exploratory variables) that can be used to predict the target (dependent or response variable). The Variable Selection node assisted in reducing the number of inputs by setting the status of the input variables that were not related to the target as rejected. Although rejected variables were passed to subsequent nodes in the process flow, these variables were not used as model inputs by successor modeling nodes. This node identified input variables, which were useful for predicting the target variables. The input status was assigned to these variables. 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 modeling nodes.

The next vertical layer of nodes consisted of the Regression (logistic for the binary target) node, Tree node, the Neural Network node. A User Defined Model was developed at one point (not in Figure 1). This node was used to develop a discriminant function. In later stages of the project the disciminant analysis modeling was discontinued because it did not performed better than the logistic regression models and more efforts were required to develop the neural network and the decision tree models.

The modeling nodes performed all of the steps required to find the most optimal model for the specific model type. Finally, the Score node (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. In practical terms it was a subset of the input data set. The predictions were accompanied by assessment statistics.

Each of the modeling nodes in Figure 1 was connected to its own Reporter node. They generated HTML reports that supported structured reporting of each modeling approach. These reports contained the process flow diagram, header information, settings, and results.

The Assessment node (Figure 1 and Figure 2) compared models and prediction diagnostics for all modeling nodes. This comparison was facilitated with a set of advanced charts for lift, and profit, return on investment (ROI), receiver operating curve (ROC), and response threshold chart[5,15,16].

A direct link between a specific modeling node and assessment node was applied to reassure the user's model selection. Alternatively, the link between the Assessment and the Score node would select the first model from the list of models, unless a manual selection was made (Figure 2a). A direct link between the decision tree node and the scoring node in Figure 2b explicitly specifies which model to score and evaluate. Non-standard and custom assessments can be applied with the SAS Code node.

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. These analyses included univariate distributions, investigation of multivariate distributions, creating scatter and box plots, displaying mosaic charts, examining correlations, and fitting explanatory models.

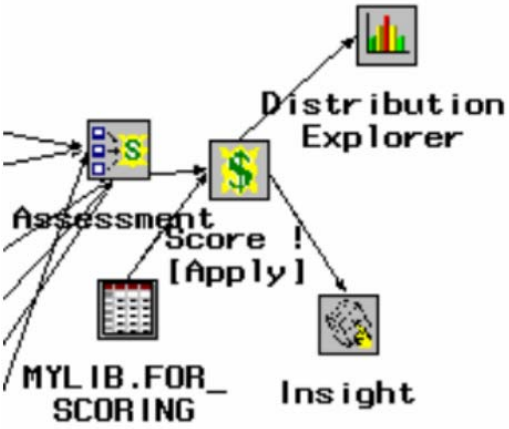


Figure 2a. The Assessment and Scoring nodes.

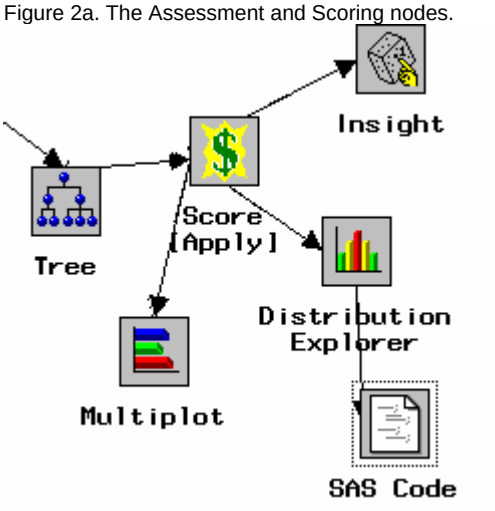


Figure 2b. The Scoring and SAS Code nodes.

**DECISION TREE MODELING**

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.

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 though 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 m3/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.

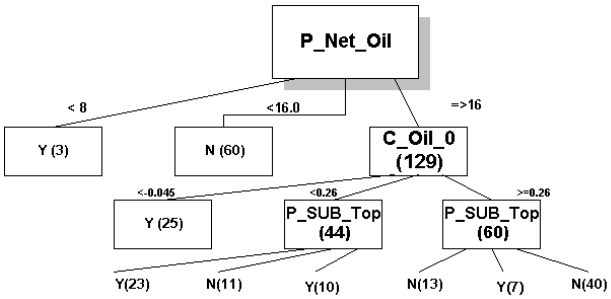


Figure 3. Tree Diagaram; Subdivision levels=3; Max Number of Splits=3.

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 bins. These leaf nodes (bins) represented the nodes that were not subdivided. Numbers in brackets show the number of observations in each leaf node (non-divided bin).

A final decision tree model in Figure 3 can be used for classifying a new well or a newly converted or treated well from the geological parameters and the instantaneous oil response (C_Oil_0). This model assigns these wells to two risk groups of good and poor producers (Y/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  < -1244.95
        AND 0.26 <= C_OIL_0
THEN
        P1      = 100.0%
        P0      = 0.0%

IF      P_SUB_TOP  < -1252.21
        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 good well
         and
        P0 is probability of poor well.
```

When a decision tree is verified and proven, 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 good (P1) and poor well (P0). With a 50% threshold cut-off value, a user's decision will be estimated from a simple formula:
if(P1 >= 50%) then GOOD else POOR.

Different assessment measures are used to select the best tree, based on the results obtained from the validation data (or test if not available). 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].

| Variable | Logworth | Groups | Label |
|---|---|---|---|
| GMC_LAG | 2.654 | 2 | |
| OMC_LAG | 2.474 | 3 | |
| RESPONSE | 2.447 | 2 | |
| C_GAS | 2.350 | 2 | |
| P_NET_OIL | 1.755 | 3 | |

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

| Variable | Worth | Groups | Label |
|---|---|---|---|
| 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 4. Competing splits for a tree with two or three branches. Splitting criterion based on Chi-square or F-test.

Larger values for both LOGWORTH and WORTH are better. The method is recursive because each set of new nodes results from splitting of the 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.


Figure 4. Example of a tree node statistics.

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.

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. Figure 4 shows an example of tree node diagnostics for each target level in percent, the corresponding count, the total count,

the overall decision level associated with a specific node. The last two rows show the expected profit for each level. The statistics is shown for the training (learning) and validation data set.

**MODEL ASSESEMENT**

Early diagnostics based on classification tables (confusion tables) indicated that decision tree models performed better than models based on logistic regression and neural networks. It 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 non-discriminatory drilling or well conversion program, a 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. This corresponds to higher oil rates.

A new well's classification in this project was based on a pre-selected cut-off applied to the estimated posterior probability. Following pseudo-code shows this logic:

If(posterior probability >= Cut-off) then Good Well
Else Poor Well

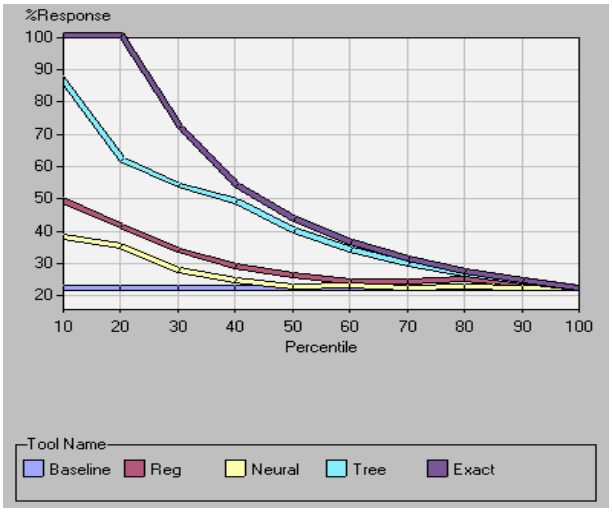


Figure 5. Cumulative response curves.

More advanced analysis and identification of the probability percentage cut-off 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 highest expected production to lowest expected production. For 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 5 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, by 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 ten percentiles along the *X* axis; the left-most percentile is the 10% of the target predicted most likely to produce. 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% or 20% or 30% of all wells than the success rate will be around 88%, 62%, and 50% respectively. This and the following charts show models being compared and superimposed with an exact model. The exact model captures all of the good wells as soon as possible (for example, if there are 30 good wells in a 100 well field, then the exact model will capture all of the wells in the first three deciles).

Figure 6 shows the percentage of good producers in each decile and this graph presents non-cmulative response curves. A baseline in this figure shows an average percentage of wells with good performance in the original sample. Thus, it is shown as a reference for any model, which was developed during this study. The presented curves show the non-cumulative response rate for the sorted deciles that correspond to percentiles from 10 to 100. The first decile (10) shows the rate (high production) for the top 10 percent of the model scores (the most likely good wells). The second decile shows the expected success rate for the second best 10 percent of the model scores, and so on. These curves allow a user to compare the model quality (success rates) in deciles (in decreasing quality bins) for different models. In particular, it shows that the decision tree model correctly predicts nearly 88 percent of the producers in the top 10% on the predicted list.

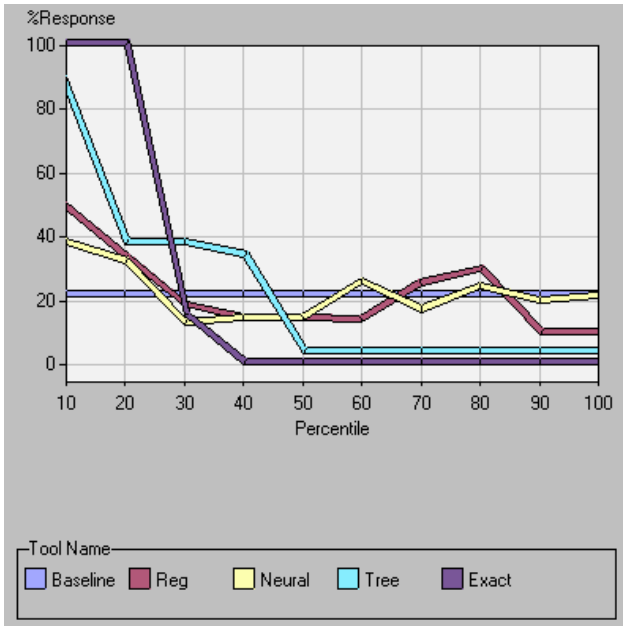


Figure 6. Non-cumulative response curves.

Figure 7 shows the cumulative lift curves, which correspond to the three models (tree, logistic regression, and neural net). A lift curve shows model's effectiveness relative to a baseline, which shows an overall (average) historical success rate (horizontal line). Non-cumulative lift curves (shown in Figure 8) enhance visual comparison of the model's performance in each decile. Figure 7 and Figure 8 show the lift curves in a relative scale, where the baseline corresponds to one (historical success rate).

The non-cumulative lift curve for the decision tree in Figure 8 shows that beyond the forth decile even the best model performed below the overall average. The non-cumulative lift curve for the tree model in Figure 8 shows nearly two to four time 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 under-performed relatively to the decision tree model. The non-cumulative lift curves in Figure 8 showed that the best model performance dropped fast from around 4 to 1.8 between the first and the third best decile. Erratic performance of the regression and neural network models between the 60 and 90 percentiles indicates their inability to account for the observed variability with the selected dimensions.

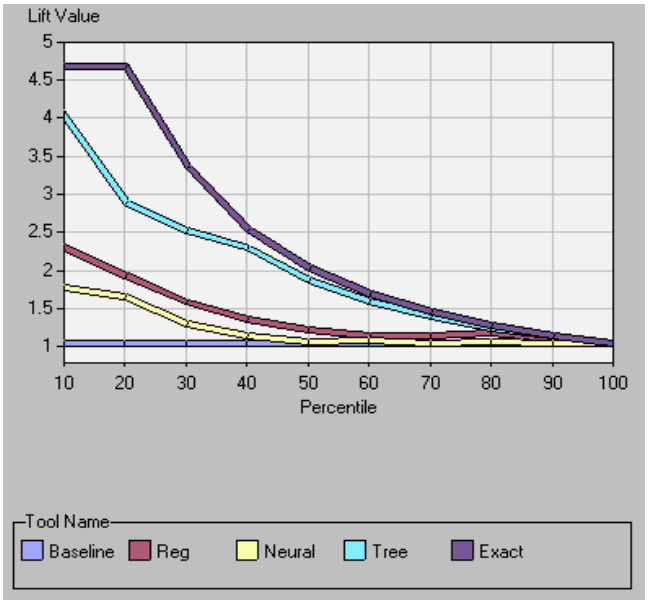


Figure 7. Cumulative lift curves.

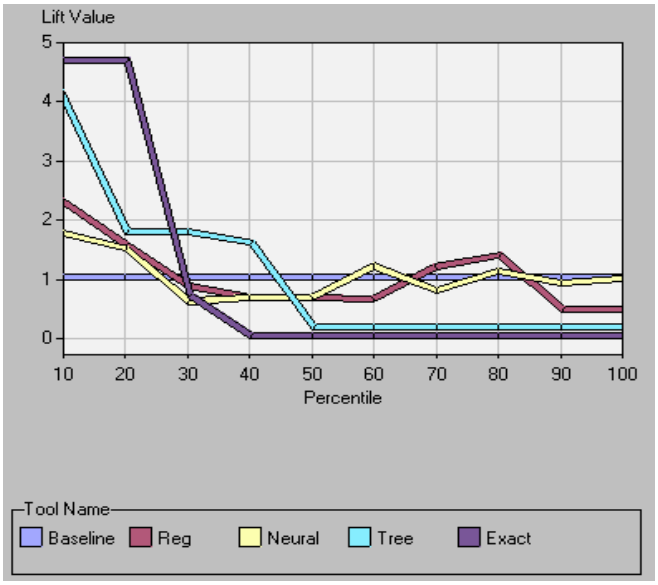


Figure 8. Non-cumulative lift curves.

As with any type of modeling, adding more variables should increase the decision tree performance. The same effect can be achieved by adding more subdivision levels (increasing the tree depth) or increasing the number of branches from a node. Finally, different node splitting criteria can make a difference in most instances. Figure 9 shows a comparison of lift curves between 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.

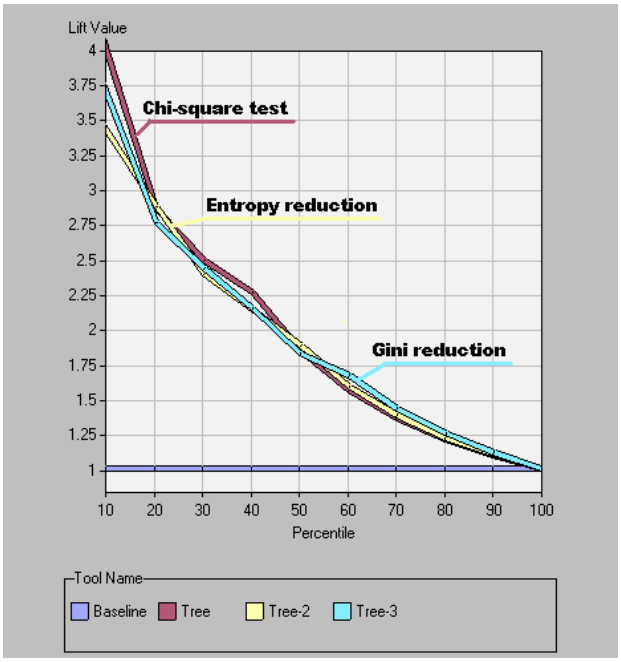


Figure 9. Cumulative lift curves for three decision trees with different node splitting criteria.

Many modeling decisions as well as the model selection depended on the misclassification rates. Figure 10 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).

A threshold-based interactive profit chart represents a higher degree in decision making. This chart enables observing the relationship between the return/profit and the threshold value for a specified profit matrix. 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 11 presents a profit matrix for these four 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 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 were tested with the model. The presented values have been changed from the original values to preserve the confidentiality of information.

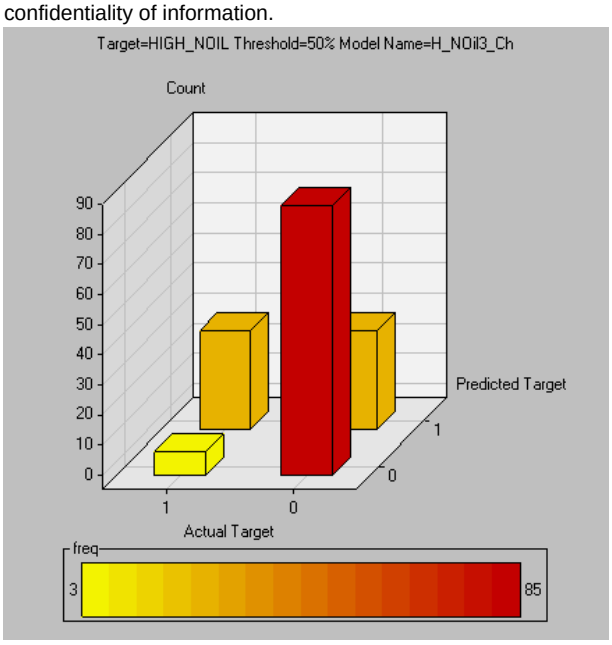


Figure 10. Threshold chart for a tree model at threshold of 50%.

The corresponding profit (return) chart in Figure 12 shows a relationship of the estimated return versus the classification threshold value (if posterior probability >= threshold, then class=1). This diagram shows that thresholds in range 5-50% should generate the highest average return. It shows that the best average return and the highest total production volumes could be achieved at the 5% threshold value. This translates to selecting most of the wells (If posterior probability >= 5% then GOOD).

The above example characterized a relatively successful waterflood implementation, where a large amount of laboratory studies were undertaken before the decisions were made.

A fine-tuning of a threshold value, which is used in a final model, is specifically important in projects with nonrandom samples (e.g. rare case sampling). During such studies, a user can uncover relationships between the predicted and the actual target values, as a function of the threshold values.

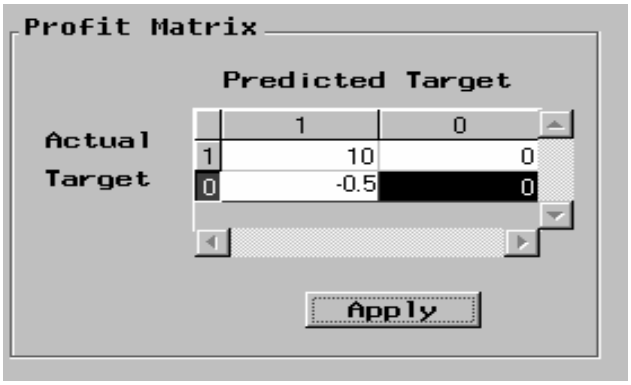


Figure 11. Profit matrix.

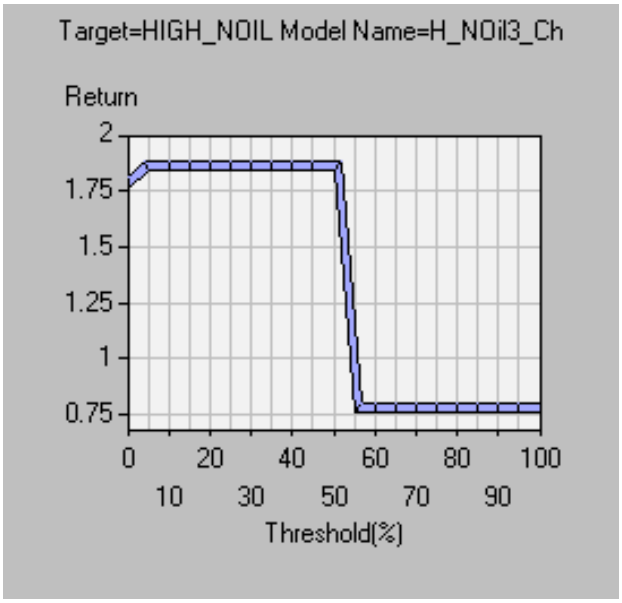


Figure 12. Return (average profit) for decision tree based on profit matrix (1/1=10M; 1/0=0; 0/1=-0.5; 0/0=0); Profit matrix from Figure 11.

Figure 13 presents similar behavior of the average return, which was based on a different profit matrix. In this example, the expected return per customer would reach a maximum at a threshold level of zero percent. Such threshold implementation corresponds to drilling of all wells. This case accounted for 'lost opportunities', which corresponded to the 1/0 case where a good well was miss-classified as a non-producer. A penalty of –8M was assigned to this miss-classification type.

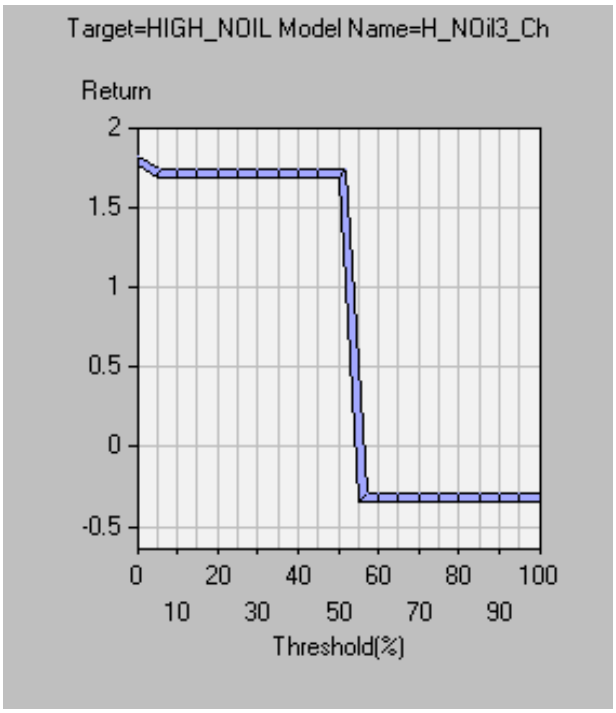


Figure 13. Return (average profit) for a decision tree based on profit matrix with a lost opportunity (1/1=10M; 1/0=-8M; 0/1=-0.5; 0/0=0).

Figure 14 shows the average profit structure for a different decision tree. In this case the best results require a much higher

threshold value. This graph shows that the low but positive expected average profit is reached for thresholds between 35% and 90%. This means that in order to maintain a high level of production and profitability the threshold value would be set to 35%. Thus, a good potential well will be defined as a well that has a posterior probability estimate greater than the threshold value of 35%.

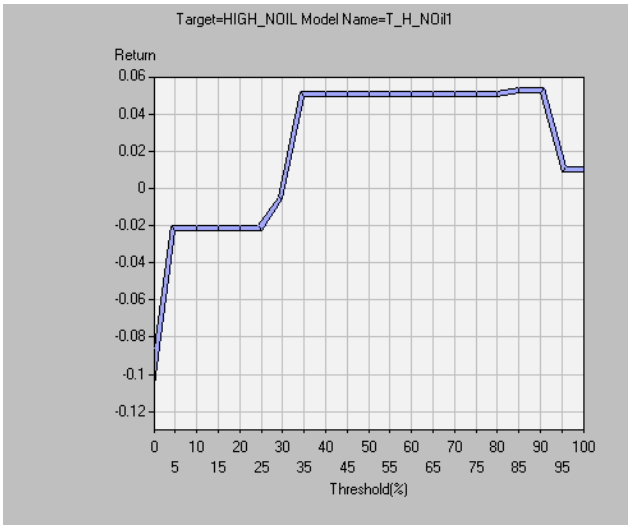


Figure 14. Return (average profit) for a decision tree based on profit matrix that recommends only the best wells.

**DECISION TREE VERSUS NEURAL NETWORKS**

Neural networks have been utilized in variety of studies with optimism fueled by the origin of this tool and from publicized successful applications. However, in this study the neural net models were not able to prove their strength[19,20,21,22].

Adding extra variables in our implementation of the neural network model (Figure 5-9) nor adding hidden layers nor adding direct links between input and output layer did not produced better models. Thus, the final model and the production system utilized a formula that was based on the decision tree model.

Furthermore, neural network development requires significant statistical analysis in order to understand the data and the process flows. Thus, most practitioners apply the stepwise regression, the backward regression, and the decision tree variable selection before applying neural network modeling. Finally, neural network models cannot be directly applied in business interpretation processes, which in some cases can eliminate the neural model from consideration. Therefore, only significantly better performance in prediction rates could justify the neural network model implementation.

**ADVANCED DIAGNOSTICS CHARTS**

The response and lift curves can be augmented with captured response curves. In previous chapter we presented definitions for cumulative and non-cumulative statistics.

For cumulative statistics, the numerator is the cumulative number of good wells in the first n deciles (between the first and the specific decile). For non-cumulative statistics, the numerator is the number of good wells in each respective decile.

For response statistics, the denominator is the cumulative number of all wells. For captured statistics, the denominator is the total number of good wells. Figure 15 and Figure 16 show the cumulative and non-cumulative statistics for three models respectively. The common criterion for all modeling and predictive tools is a comparison of the expected to actual profits or losses obtained from model results. This criterion enables us to make cross-model comparisons and assessments, independent of all other factors (such as sample size, modeling node, and so on).

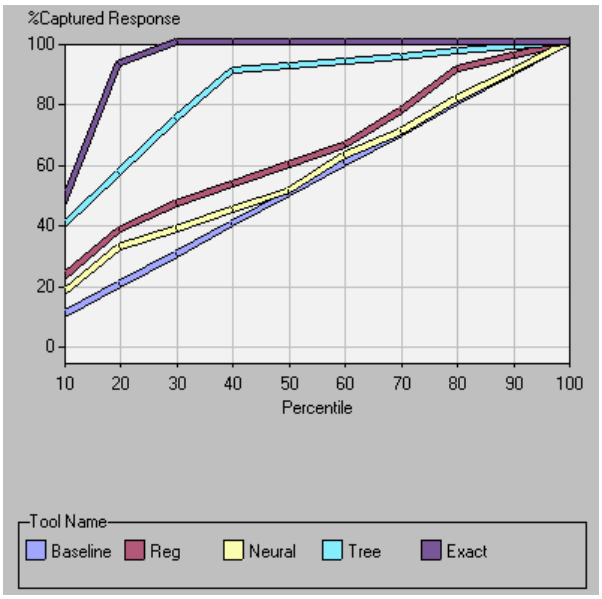


Figure 15. Cumulative captured response curves.

Average expected profit curves are presented in Figure 17 and Figure 18, while the ROI (return on investment) curves are presented in Figure 19 and Figure 20. Similarly to other diagrams (charts) they were calculated based on the verification set and the expected outcome came from the profit table (see Table 1). The return on investment (ROI) chart displays the cumulative or non-cumulative ROI for each decile of observations in the score data set. The return on investment is the ratio of actual profits to costs, expressed as a percentage.

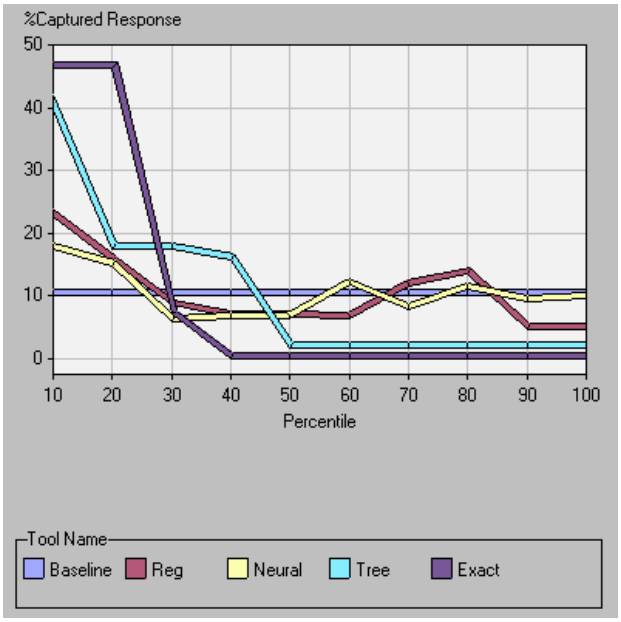


Figure 16. Non-cumulative capture response curves.

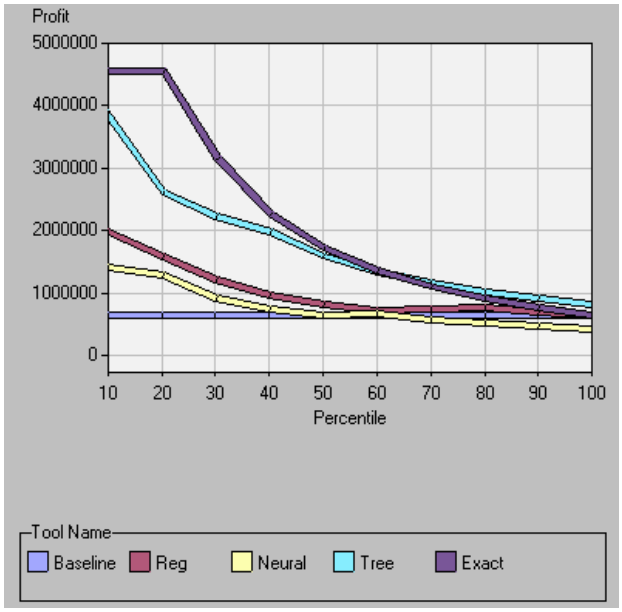


Figure 17. Cumulative profit curves.

Classification charts (Figure 10) display the agreement between the predicted and actual target variable values for non-interval target variables. If the predictive model is useful, then agreement will be good, and the tallest bars will appear on the main diagonal of the chart. If the predictive model is not useful, then agreement will be poor, and the bars will be roughly proportional to the product of the row sum and column sum throughout the chart. We can display either counts or percentages of the predicted and actual target values. Percentages are especially useful when at least one of the target levels is rare. Figure 21 shows a confusion table with definitions corresponding to binary outcomes.

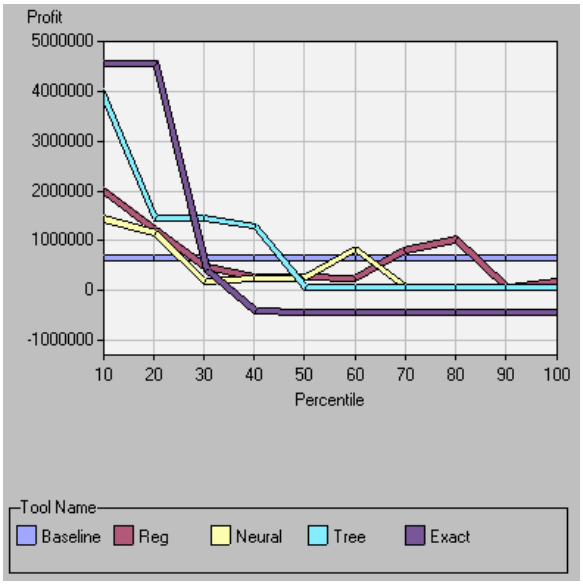


Figure 18. Non-cumulative profit curves.

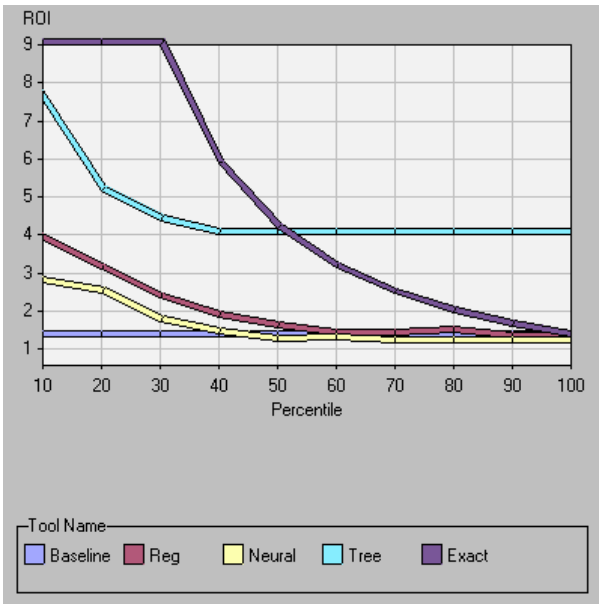


Figure 19. Cumulative ROI curves.

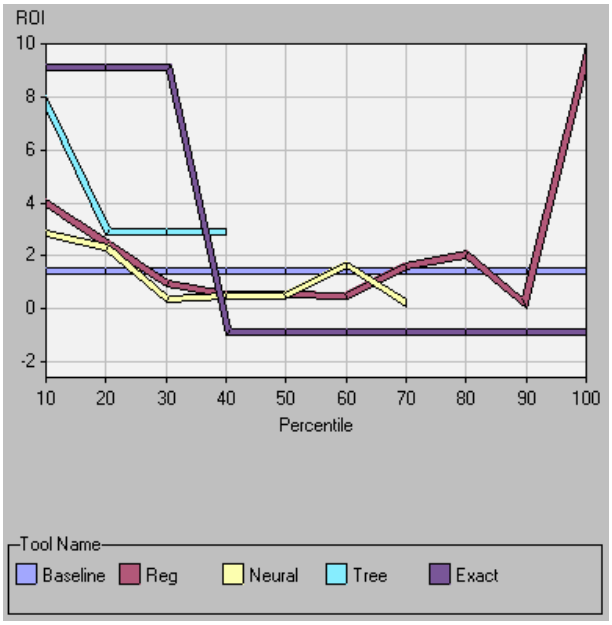


Figure 20. Non-cumulative ROI curves.

A measure of discrimination of the ROC curves is the area between the specific curve and the diagonal line. This diagonal line represents a random choice between positive and negative events (no discrimination or no skill line). The ROC curve comparison confirms the strength and superiority of the decision tree model developed in this project.

Prior probabilities or frequencies effect both the confusion charts and corresponding confusion tables. Thus, any diagnostics that is based on 'accuracy' is not appropriate to compare models evaluated based on different samples.

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Table of actual by predict | | |
|---|---|---|---|
| | predict | | |
| actual | 0 | 1 | Total |
| 0 | TN<br>Specificity | FP | |
| 1 | FN | TP<br>Sensitivity | |
| Total | | | |

Figure 21. Confusion Table – Definitions

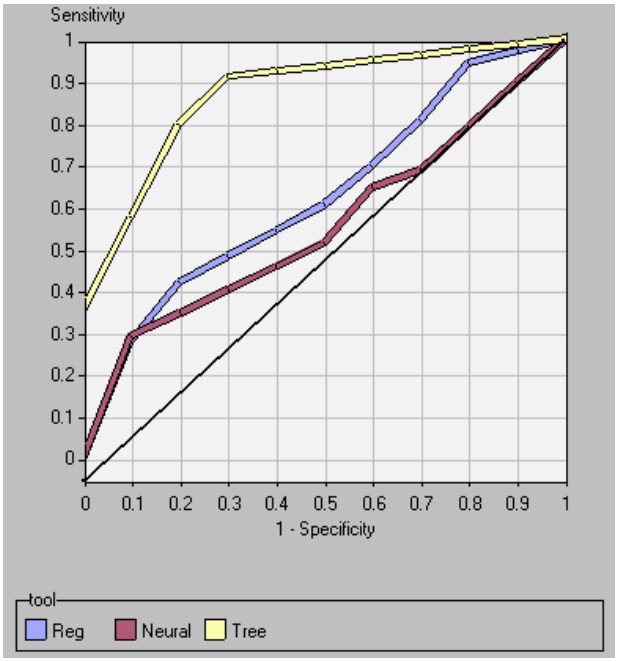


Figure 22. ROC curves.

A true measure of discrimination between positive and negative populations is represented by Receiver Operating Characteristics (ROC) curves, which are based on a cross-plot of sensitivity as a function of (1-Specificity). This definition is based on actual (row) counts where:

Sensitivity is the accuracy of predicting positive events:

*Sensitivity=(True Positive)/(Total Actual Positive)*

and

specificity is the accuracy of predicting non-events or negatives:

*Specificity=(True Negative)/(Total Actual Negative)*

Sensitivity is often referred as the *Hit Rate* and *1-specificity* is called *False Alarm Rate*. These two descriptors are often called *True Positive (TP)* and *False Positive (FP)*. Furthermore, both of them are independent of the prior probability of good and poor wells. Thus, ROC curve in Figure 22 does not depend on the prevalence of the target outcome in the actual population and provides target outcome predictability that is independent from the prevalence and decision threshold effects.

A comparison of the prediction accuracy is be presented with the Response Threshold charts in Figure 23.

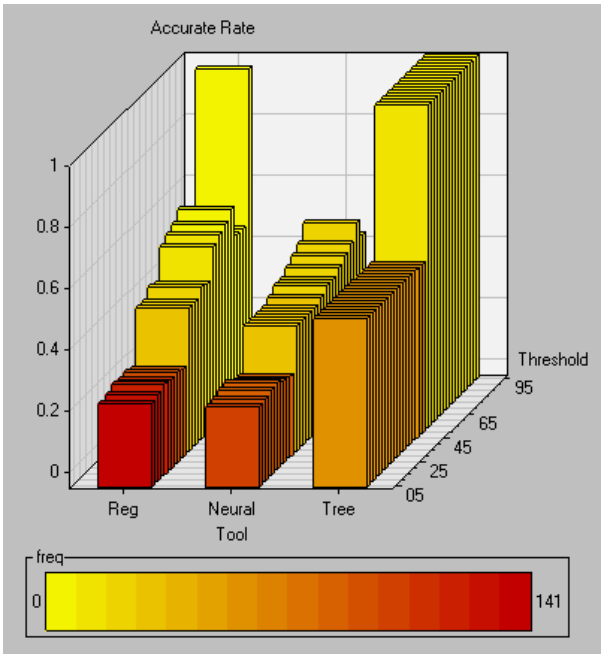


Figure 23. Response Threshold Charts.

The response threshold charts display the prediction accuracy of the target level across a range of threshold values. This diagnostics is based on predicted counts and column percentages (see confusion table). These column percentages are often called Positive Predictive Value (PPV) and Negative Predictive Value (NPV), which correspond to Probability of '1' when predicting '1' and Probability of '0' when predicting '0' respectively.

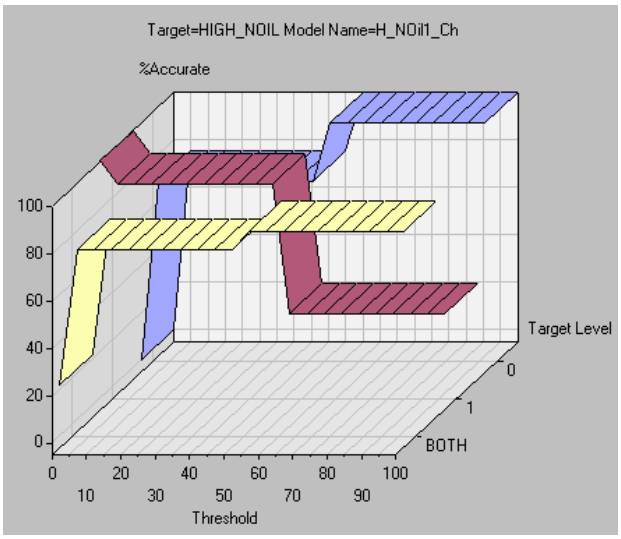


Figure 24. Correct Classification Chart.

A correct classification of both levels for the decision tree model is shown in Figure 24.

Charts presented in this chapter were generated from a series of calculation for thresholds ranging from zero to one. Example two confusion tables for two selected thresholds are presented in Figure 25 and Figure 26.

**thresh=10**

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Table of actual by predict | | |
|---|---|---|---|
| | predict | | |
| actual | 0 | 1 | Total |
| 0 | 85<br>59.03<br>75.22<br>96.59 | 28<br>19.44<br>24.78<br>50.00 | 113<br>78.47 |
| 1 | 3<br>2.08<br>9.68<br>3.41 | 28<br>19.44<br>90.32<br>50.00 | 31<br>21.53 |
| Total | 88<br>61.11 | 56<br>38.89 | 144<br>100.00 |

Figure 25. Confusion table for Threshold=10.

**thresh=75**

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Table of actual by predict | | |
|---|---|---|---|
| | predict | | |
| actual | 0 | 1 | Total |
| 0 | 113<br>78.47<br>100.00<br>84.96 | 0<br>0.00<br>0.00<br>0.00 | 113<br>78.47 |
| 1 | 20<br>13.89<br>64.52<br>15.04 | 11<br>7.64<br>35.48<br>100.00 | 31<br>21.53 |
| Total | 133<br>92.36 | 11<br>7.64 | 144<br>100.00 |

Figure 26. Confusion table for Threshold=75.

A cross lift and profit charts are the same as other charts, except that they plot the lift or profit on two or more partitions of the data. In Figure 27 we use cross lift charts to compare the lift for the decision tree model obtained from the training data to the lift for the validation data, and to the testing data (40%, 30%, and 30% of the total observations in these sets). Differences in the performance are expected. However, they might indicate potential problems with the model and its robustness (generalization) when applied to unknown data sets. However, a cross profit chart in Figure 28 shows profit differences between different partitions. It indicates much better performance of the testing set comparing to the learning and validation sets.

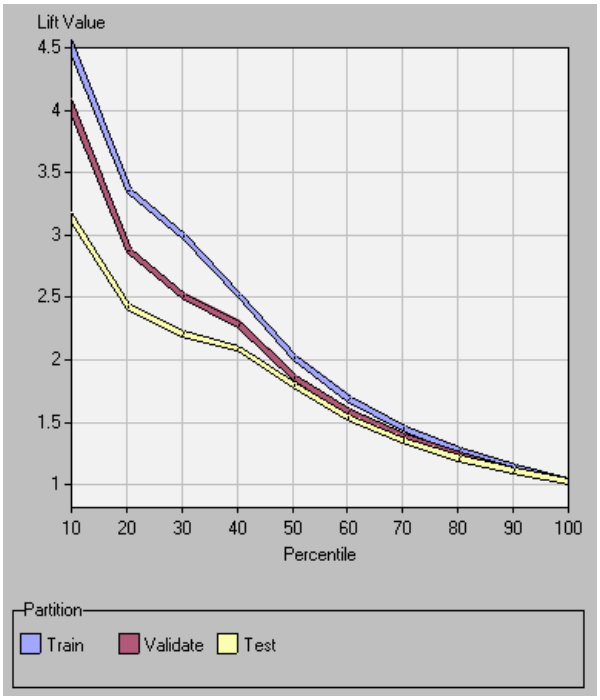


Figure 27. Cumulative Cross-Lift curves for the decision tree model.

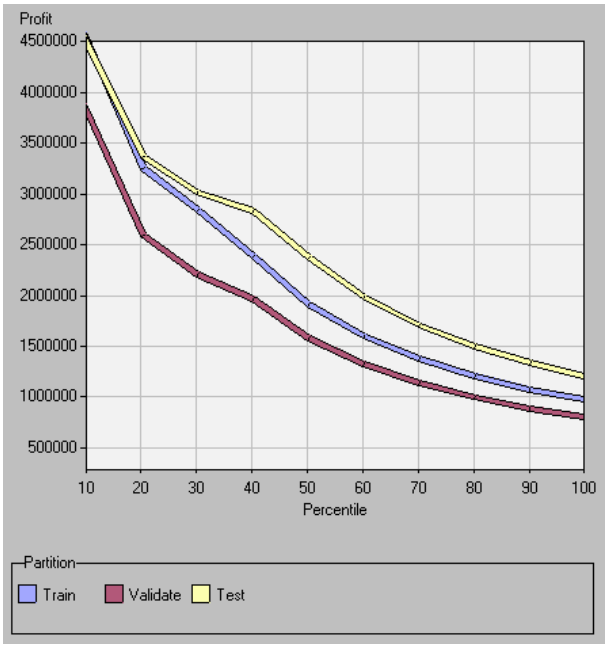


Figure 28. Cumulative cross profit charts for the decision tree models.

## ADVANCED MODELS

### NEW DATA SETS

In next stage of development, two additional data sets were obtained and merged with the project data base that contained the waterflood responses and basic geology. The first set contained core data. After loading, this data went though intensive process of cleansing, normalization, up scaling, and missing replacement. The core data contained:

- formation sample (SampleForms),
- lithology (Lithology)
- sample number (S_No)
- sample top (Sample_Top)
- sample base (Sample_Base)
- thickness in 'm' (Thickness_M)
- three permability meassurements (Kmax_mD, Kvert_mD, K90_mD)
- porosity (Porosity)
- grain density (Grain_Density)
- bulk denisty (Bulk_Density)
- residual oil sateration (Rsat_Oil_Ratio)
- residual water saturation (Rsat_Water_Ratio)
- sample formation (SampleForms).

The second data set contained names, tops, and perforation lengths in each formation. The formation and perforation data were calculated, transformed, and classified for the most common formations ELRL, PKSK, BNFF, SHND, and OTHR (remaining) formations. Several new parameters and indicators were developed to be testing in the modeling development.

Finally, the whole merged data based was transposed into a set with parameters that characterized both wells in each producer-injector pair. Furthermore, a new parameter was generated that characterized the producer-injector connectivity based on the producer-injector formations where perforations were shot.

### RECENT MODEL DEVELOPMENT

The latest model development was directed in testing different model techniques and optimizing the decision tree models. The latest project diagram is presented in Figure 29.

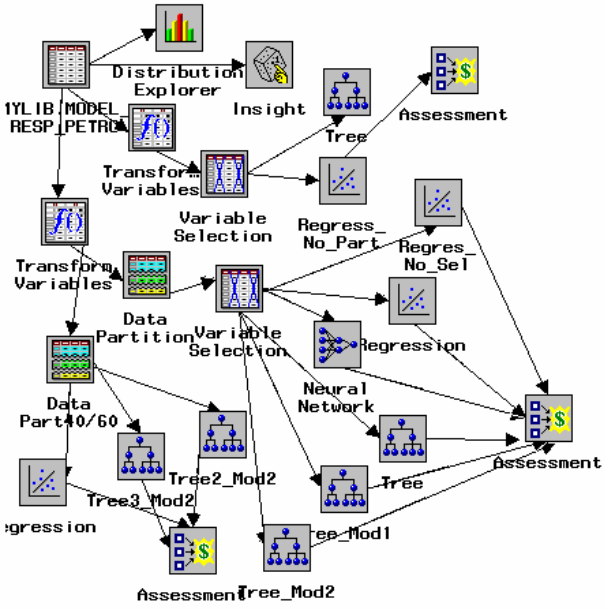


Figure 29. The latest project diagram.

The tree models have been outperforming the neural net and regression models. In addition, over-sampling was tested. In latest case the learning and validation partition contained 60% and 40% of the total number of observations from the project database.

A set of cumulative lift curves together with the exact lift curve Is presented in Figure 30.

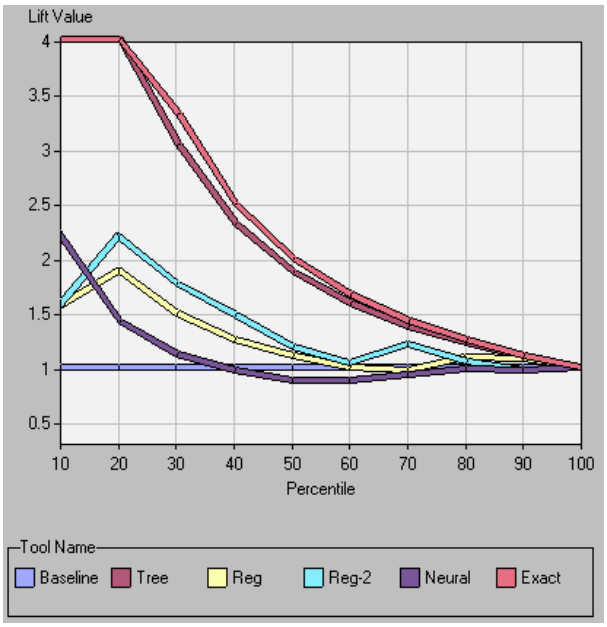


Figure 30. Cumulative lift curves.

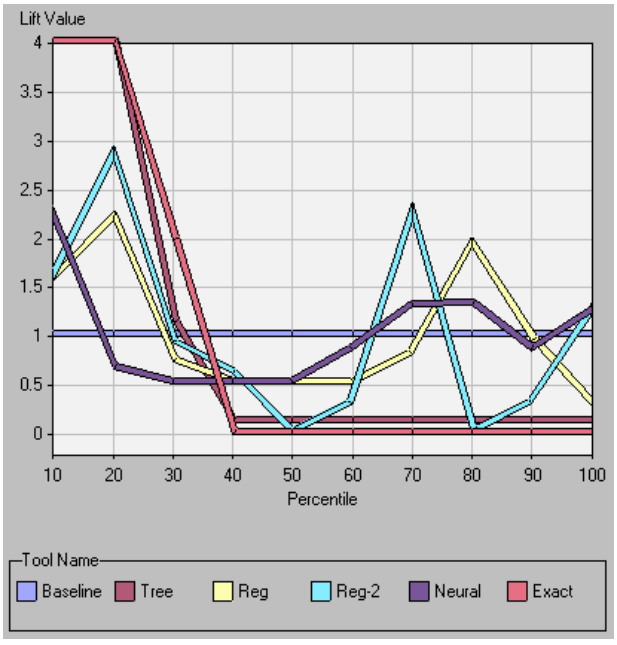


Figure 31. Non-cumulative lift curves.

The above two sets of lift curves show near perfect performance of the decision tree model. The actual lift curve nearly match the exact lift curve. In particular, diagrams show that the first two deciles show no misclassifications. At the same time the cross

lift curves in Figure 32 confirm the solid model performance. The statistics for the testing set is worst than for the learning and validations sets. However, it is still overwhelmingly better than in the development without the core and completion data.

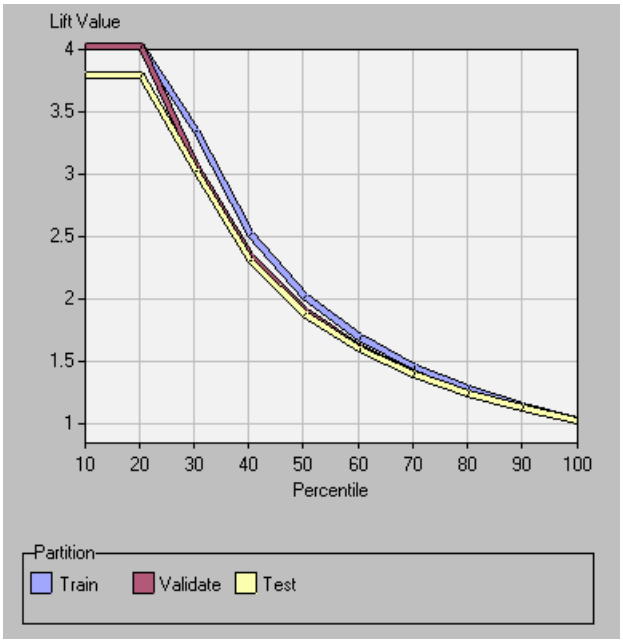


Figure 32. Cross lift curves for the decision tree model.

Such a good performance always triggers a question: 'Are we guilty of model over-fitting?' In order to answer this question a set of models were generated with the over-sampled learning and verification sets (60, 40, 0 percentages for learning, verification, and testing set accordingly). Lift curves for a regression model and two tree models (see Regression, Tree2_Mod2, and Tree3_Mod2 models in the lower left hand side of the project diagram in Figure 29) are presented in Figure 33 and Figure 34.

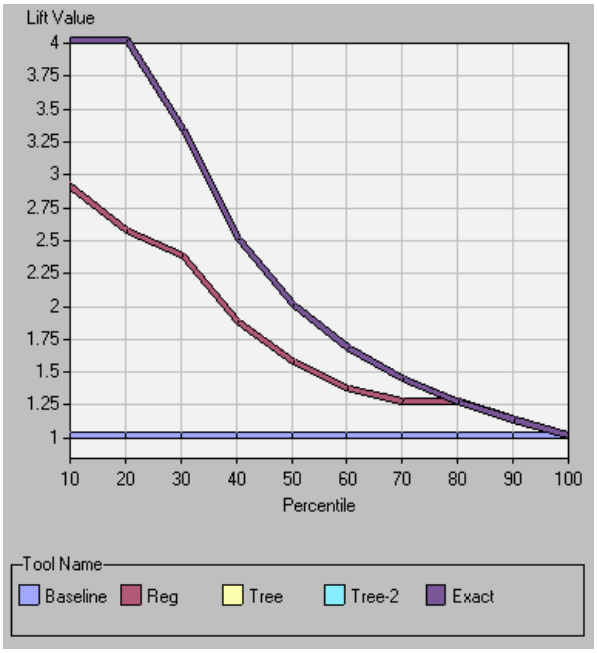


Figure 33. Cumulative lift curves and the exact curve. Over-sampling.

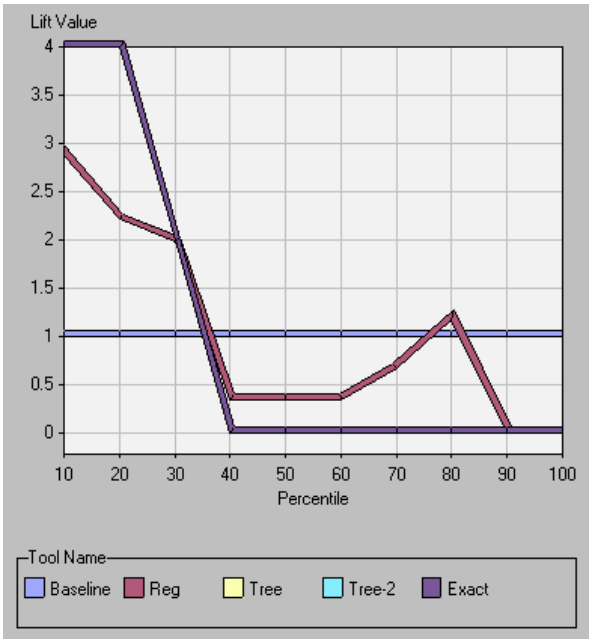


Figure 34. Non-cumulative lift curves and the exact curve. Over-sampling.

Lift curves for the tree models correspond to the exact lift curve, which translates to the 'perfect fit'. This requires further studying and is being research. Over-sampling seems to have a significant effect on the regression model, which is stronger than previously developed model and performs consistently across all deciles for both partitions (learning and verification – see Figuer 35). At this stage models were developed without considering the profit matrix and constant costs vector.

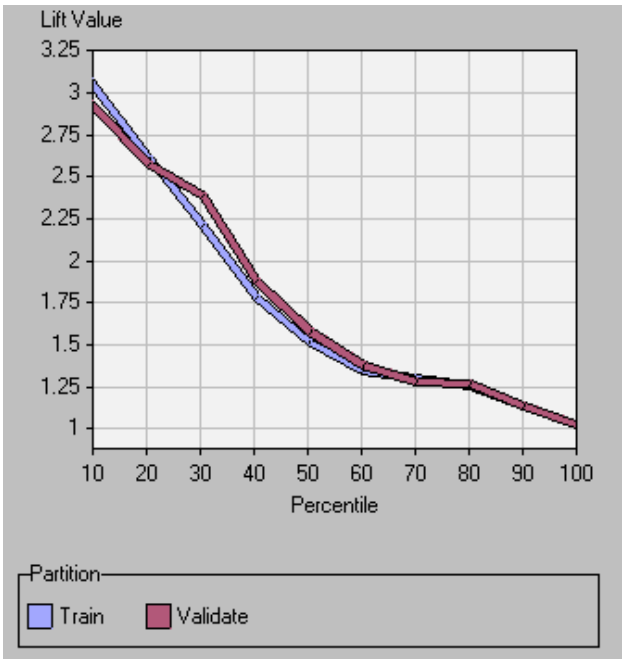


Figure 35. Cumulative cross lift curves for the regression model.

## TOOLS AND IMPLEMENTATION

### TOOLS AND MODELING DECISIONS

The case described in this paper is not complete. Hopefully it provides enough arguments for scientific based modeling and methodology selection, which should be driven by results and their diagnostics.

Process modeling and data mining are not for amateurs. The use and abuse of the advanced mining tools can lead to disastrous results. Even, a complete rookie can create a model and generate predictions using some of the tools available on the market. Advanced modeling relies heavily on data manipulation, normalization, transformations, and selection of the most important variables, model type, and variety of different options when developing and optimizing selected models[23].

The user has to look for diagnostics and interpret results from an engineering point of view. Finally, erroneous data manipulation and transformation can render the modeling part useless.

### COMPUTER DECISION SYSTEMS IMPLEMENTATION

Production implementation of decision support systems introduces an extra layer of complexity. The system requirements in cases where humans are replaced by intelligent systems expand to lesser-known regions of the systems development. User interfaces, data gathering and simple reports are not enough. Typically, when models are developed, the implementation follows without delays. However, after an initial period, the system sponsors, technical staff, and support people start to realize that the model implementation was simply a good start along a bumpy road.

In a manual process, a series of rules, user's perceptions, and unwritten rules are applied to each decision. Mistakes are made, though their impact is minimized because in manual processes these errors are inconsistent. For example, different users apply different variations of rules and their knowledge changes in time. Thus, even a weak set of rules does not have to generate wide spread problems.

On the contrary, computer based models make decisions for many wells in a consistent way. These systems require their models to be replaced or modified when the maintenance is scheduled. What happens if there are drastic changes in the environment and the maintenance is not applied? The answer is simple. The non-maintained system happily generates more or less useless predictions, which are applied as designed. Furthermore, even the best models make mistakes, and an array of non-believers and the 'old guard' team will find plenty of examples 'proving' that the system is useless.

Thus, decision systems should be equipped with a reporting and visualization tools based on data dimensions, multidimensional reporting, graphics, statistical diagnostics, and the system performance diagnostics. In a perfect world any well data together with the predicted and the actual results should be available to a user in an interactive EIS/GIS system.

## CONCLUSIONS

This study proved that integrated geological and waterflood response parameters allow for the prediction of oil production during enhanced recovery processes. Different model types were built, which included decision trees, regression models, and neural networks. These predictive models were developed for two types of target variables. The first target variable represented the normalized oil production, which was characterized by a well production relative to the whole field's production. The second target variable was characterized by a high production binary indicator with two levels (0/1 or Good/Poor).

The final model was build for the binary indicator. It was selected based on the oil production cut-off. Logistic regression, neural net, and decision tree models were developed and compared based on their diagnostics. We selected the decision tree model due to the best performance. An advantage of the decision tree model over other types of models was that it could produce models that represented interpretable English rules or logic statements. For example, "If netpay is greater than 5m and the lag zero gas response is negative, then oil production will be in the top 25% of the best production with probability of 80%".

Statistical diagnostics based on the model verification process proved that selecting wells based on models that use geological and fluid communication parameters resulted in a success rate two to four times better than by traditional methods. The final decision tree model showed 'perfect' performance without miss-classification. This requires detailed review and further research.

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

Model based computer decision systems require appropriate model selection, model diagnostics, model maintenance schedules, and information visualization. Diagnostics of model's performance must be carefully designed and implemented. Furthermore, simplicity, interpretation ability, maintenance requirements, and stability of models should influence both the modeling and the development approach.

## REFERENCES


1. John C. Davis, "Statistics and Data Analysis in Geology", John Wiley & Sons, 2nd Edition, 1986, ISBN 0-471-08079-9.
2. SAS Institute Inc, Logistic Regression Examples, Version 6, First Edition, ISBN 1-55544-674-4.
3. Leon Fedenczuk, Paul Pedersen, Marc Marshall, Analyzing Waterflood Responses for Pekisko B., Journal of Canadian Petroleum Technology, vol. 40, No. 6, June 2001, pp. 29-35.
4. SAS/STAT User's Guide, Version 6, Fourth Edition, ISBN 1-55544-376-1.
5. Data Mining Using Enterprise Miner Software: A Case Study Approach, First Edition, ISBN-58025-641-4, February 2000.
6. Leon Fedenczuk, Paul Pedersen, Marc Marshall, The Petroleum Society of CIM – 50th, Paper 99-46, Annual Technical Meeting, June 14-18, 1999.
7. L. Fedenczuk, K. Hoffmann, Surveying and Analyzing Injection Responses for Patterns with Horizontal Wells, International Conference on Horizontal Well Technology, November 1-4, 1998, paper No. SPE 50430.
8. Mark L. Berenson and David M. Levine, "Basic Business Statistics Concepts and Applications", Prentice-Hall, Inc., Englwood Clifs, New Jersey 07632, 1983, ISBN 0-13-057620-4.
9. L. Fedenczuk, K. Hoffmann, "Data Integration and Analysis for Optimal Field Development", The Petroleum Society of CIM - 48 th Annual Technical Meeting, June 8-11, 1997, paper 97-45.
10. Michael J.A. Berry and Gordon Linoff, Data Mining Techniques, Data Marketing, Sales, and Customer Support, John Wiley & Sons, Inc., ISBN 0-471-17980-9, 1997.
11. Bishop, C. M. (1995**),** Neural Networks for Pattern Recognition, New York: Oxford University Press.
12. Bigus, J. P. (1996), Data Mining with Neural Networks: Solving Business Problems - from Application Development to Decision Support, New York: McGraw-Hill.
13. L. Breiman, J.H. Friedman, R. A. Olsen, and C. J. Stone. Classification and Regression Trees (1984), Pacific Grove: Wadsworth.
14. Chidanand Apté and Sholom M. Weiss, Data mining with decision trees and decision rules, Future Generation Computer Systems, Vol. 13, pp. 197-210, 1997.
15. Charles E. Metz, Basic Principles of ROC Analysis, Seminars in Nuclear Medicine, Vol. VIII, No. 4 (October),

1978, pp.283-298.
16. James A. Hanley, Barbara J. McNeil, The meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve, Radiology 143, pp29-36, April 1982.
17. A. Feelders, Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation, Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery (PKDD-99), LNAI, Vol. 1704, pp. 329-334, Springer, September 15-18 1999.
18. Manish Mehta and Jorma Rissanen and Rakesh Agrawal, MDL-Based Decision Tree Pruning, Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95), pp. 216-221, August 1995.
19. L. Zhang and B. Zhang, Neural Network Based Classifiers for a Vast Amount of Data, Lecture Notes in Computer Science, Vol. 1574, pp. 238-246, 1999.
20. S. M. Monzurur Rahman and Xinghuo Yu and Geoff Martin, 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.
21. C. Deng and F. Xiong, Neural Method for Detection of Complex Patterns in Databases, Lecture Notes in Computer Science, Vol. 1574, pp. 258-262, 1999.
22. Carsten Jacobsen and Uwe Zscherpel and Petra Perner, A Comparison between Neural Networks and Decision Trees, Proc. 1st Int. Work. Machine Learning and Data Mining in Pattern Recognition, MLDM, Lecture Notes in Artificial Intelligence, LNAI, Number 1715, Springer-Verlag, September 1999.
23. M. Terabe and O. Katai and T. Sawaragi and T. Washio, A Data Pre-processing Method Using Association Rules of Attributes for Improving Decision Tree, Lecture Notes in Computer Science, Vol. 1574, pp. 143-147, 1999.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged.
Please contact Leon Fedenczuk at:

Gambit Consulting Ltd.
268 Silvercrest Dr. N.W.
Calgary, Alberta, T3B 3A4 Canada
Tel: (403) 288-6754
Email: leon@gambitconsulting.com