Oil Data Analysis

James Phillips

Edward Cripps
School of Mathematics & Statistics

Melinda Hodkiewicz
School of Mechanical & Chemical Engineering

Martin Weston & Robert Taylor
CEED Client: WesTrac Pty Ltd

Abstract

WesTrac Pty Ltd (WesTrac) is one of the largest Caterpillar® dealers in the world that provides condition monitoring services to clients in the form of oil analysis and maintenance suggestions. This project uses diagnostic oil data to develop logistic regression (LR) models to classify oil samples taken from different machinery compartments into categories relating to the condition of the oil. Such classification is currently undertaken manually and as a result is potentially subject to interpretation bias and human error. Models developed to help screen 'A samples' (those with no contamination) correctly classified 85-96% of the samples tested. Multinomial LR models extend work done Ratnam (2011) by classifying the oil data into the four categories (A, B, C, X) currently used at WesTrac. An investigation is also made into the key factors that influence the oil sample classification and of most significance, the effect of previous compartment history on future classification is outlined

1. Introduction

Empirical evidence indicates oil sampling is an effective condition monitoring technique used in condition based maintenance, which allows the identification of compartment wear before failure occurs (Williams et al, 1995). Such maintenance requires an accurate and consistent measurement of deterioration in order to be effective. The Equipment Management Centre (EMC) at WesTrac currently evaluates over 400,000 oil samples each year from Caterpillar machinery by processing information gathered from a range of different tests in WesTrac's oil laboratory. Test results confirm the condition of the lubricating oil and provide important information on the compartments' condition. The interpretation separates the overall compartment and oil condition into the following classifications.

A	Oil properties are within acceptable limits and operation can continue as usual
B	Certain results are outside acceptable ranges - minor problems with the machinery
C	Unsatisfactory results are present - significant problem with the compartment and lubricant properties
X	Clear contamination needing immediate diagnostic and corrective action to prevent possible failure

Table 1: Oil sample classification levels

WesTrac has identified an opportunity to improve its current process for the classification of oil sample information. Much of this analysis is currently undertaken manually and as a result is potentially subject to interpretation bias and human error. This project uses logistic regression (LR), a type of regression analysis for predicting the outcome of a categorical

variable, to classify oil data. The classification models will enable WesTrac to improve service to its clients and assist the predictive maintenance of Caterpillar machinery by simplifying and speeding up the classification process. It will also support a move to a more automated interpretation system allowing interpreters to spend more time on those samples most critical to machinery failure.

1.1 Background

The difficulty posed by prediction and classification problems has resulted in a large number of problem-solving techniques, each with different benefits and disadvantages (Kim, 2009). LR is advantageous as it does not rely on assumptions of normality for the predictor variables or errors (unlike discriminant analysis), and it allows the selection effect to vary nonlinearly (Janzen et al, 1998). Although the use of LR techniques in condition monitoring applications is increasing, previous work has mostly seen the classification of binary data using a small number of predictor variables. Model-building techniques used in this project are similar to the methods in Jardine et al (2005) & Aguilera et al (2005) for dealing with problems of multicollinearity, and Hosmer et al (2000) for the transformation of continuous variables.

More complex multinomial LR models have also been developed and tested which classify oil samples into the four categories in use at WesTrac, thus exploring new ground in condition based maintenance. Multinomial LR has been extensively applied in medicine, ecology and economics but its use in reliability has been limited to date. This builds on the work done by 2011 CEED student Mithran Ratnam who investigated the use of neural networks as a decision support tool to help screen 'A samples' out of the manual analysis process. The software NeuroShell Classifier was adopted and approximately 90% of the 'A samples' of each compartment tested were correctly classified.

2. Process

2.1 Data Collection

LR models have been applied to a number of machinery compartments from a fleet of 60 trucks at a mine site known for providing accurate data. A focus has been on engine oil due to the complex nature of engines and the numerous condition indicators presented. Once the data is exported from *Oil Commander*, WesTrac's online database, as a .csv file, errors, irrelevant variables and missing data points are corrected and cleaned. Statistical analysis has been performed using 'R'.

2.2 LR Model Development

Figure 1 highlights the process of the model development, which is based on methods presented by Hosmer and Lemeshow (2000). The process continues until it appears that all of the important variables are included and those excluded are scientifically and or statistically unimportant. Once the regression model and its covariates are decided a comparison is made with a number of statistical model development methods. In this work the results of both model selection techniques agreed.

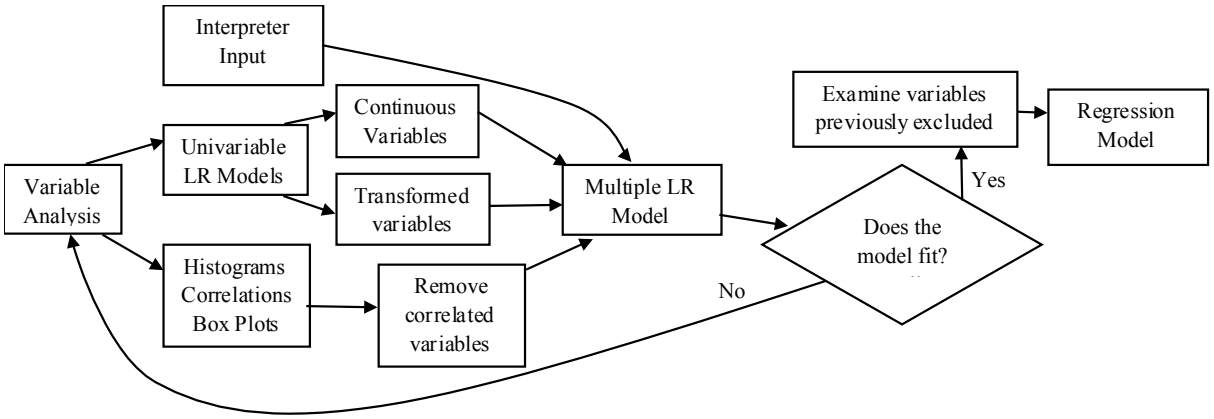

Figure 1: Model Development Process

2.3 LR Model Validation

Since predicting the classification outcome of an oil sample is important for WesTrac, the LR models are validated through their predictive performance. The prediction of each sample tested was defined as the outcome with the highest probability, i.e. in the binary case the outcome with a probability greater than 0.5. To estimate the predictive performance 80% of a data set is selected randomly to fit the model and the fitted model is used to predict the remaining 20% of the data. To obtain a better idea of the distribution of the results, the process is repeated 100 times. Thus allowing a more accurate prediction as to how the model will perform when classifying oil samples not seen before.

3. Results and Discussion

3.1 Binary Classification Models

Over the numerous data sets tested for the binary case, 85- 96% of all samples were classified correctly, on par with the neural networks approach implemented by Ratnam (2011). Additionally, logistic regression models allow variables to be examined and related to possible sources of machinery failure. During the model development it is possible to determine the key variables that are related to the classification outcome using likelihood ratio tests and via the examination of variable Wald statistics. Iron clearly showed the highest amount of influence on the classification outcome, a consequence of its affiliation with the wear of cylinder linings. Table 2 highlights the variables included in the final model for classifying engine samples in the binary case.

Variable	Coeff. β	Odds	Variable	Coeff. β	Odds
Fe	-0.259	0.772	oilhours	0.005	1.005
Pb	-0.349	0.705	prevB	-1.483	0.227
binCu	-0.738	0.478	prevC	-2.002	0.135
binSi	-0.808	0.446	prevX	-1.973	0.139
binNa	-0.689	0.502	OXI	-0.107	0.899

Table 2: Binary LR models coefficients and odds ratios

Many authors (Hosmer et al, 2000, Aguilera et al, 2005) describe how LR models become unstable due to high dependence among predictor variables (multicollinearity). As a result of multicollinearity the interpretation of the relationship between the classification outcome and the explanatory variables in terms of odds ratios may be incorrect. Such problems exist for the oil data presented and as a result this effects the inclusion of critical variables such as soot.

Principle component analysis (PCA) can be applied in this instance to reduce such problems of multicollinearity and dimensionality (Jardine et al, 2006 & Aguilera et al, 2005) and is currently being explored to further enhance the classification accuracy of the models.

When problems with multicollinearity are non-existent the variable coefficients are illustrative in explaining their relationship with the oil sample classification outcome. In LR models a unit increase in an explanatory variable is associated with a change in the odds by a factor of exp(β). For example the odds ratio for lead is 0.705, which implies that a 1 unit (ppm) increase in lead decreases the odds of an oil sample being classified as 'A' by 0.705. The odds ratio for oil hours of 1.005 may appear counter intuitive as this suggests with increased oil hours there is higher odds of the sample being classified 'A'. However, the variable indicates the interaction with the wear particles such as iron and lead. It accounts for the fact that an iron level of 20ppm for a 250hr sample should have lower odds of being classified as A than the same iron level for a 500hr sample.

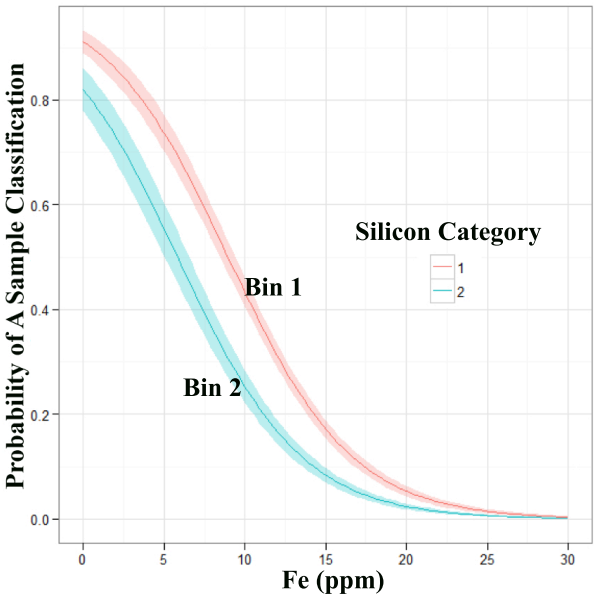

Figure 2: Effects of Fe & Si on the probability of the occurrence of an 'A sample'

Figure 2 highlights similar information which is a plot of the probability of a sample being an 'A sample' for a given level of iron and silicon. Silicon is transformed into two groups, below and above the median. Similar figures can be developed for different variables that can be used by WesTrac to provide more interpretable messages to clients, and for in house training.

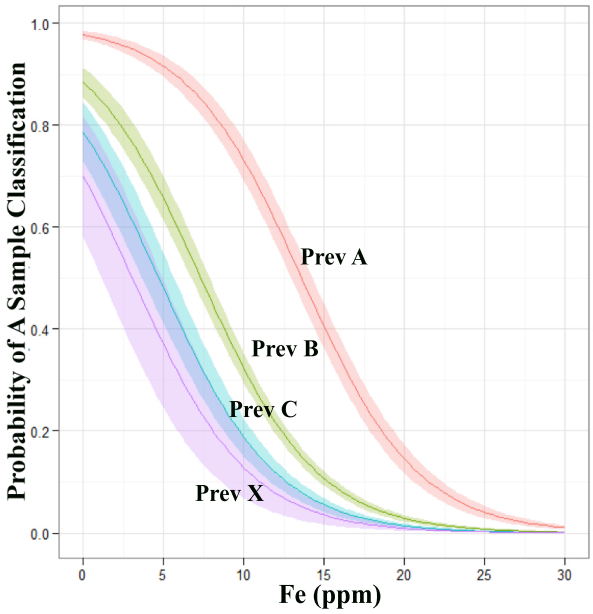

Figure 3: Effects of previous history & Fe on the probability of the occurrence of an 'A sample'

3.2 Oil Compartment History

The variable 'prev' was introduced to reflect the importance of previous history to future engine performance. 'Prev' represents the classification that the same piece of machinery and compartment received the last time an oil sample was classified. Figure 3 highlights the effect of this on the classification of each sample. For example if an oil sample has an iron level of 10ppm and the compartment was last time classified 'A', then the estimated probability of it being classified as an 'A sample' is 0.75. For the same level of iron, if the compartment was last time classified 'X' then the estimated probability of it being an 'A sample' is 0.1.

This highlights the importance of machinery past history on future performance and classification. From a classification point of view it suggests knowing the history of machinery when interpreting data would help improve classification accuracy. In relation to this, LR models for the analysis of time dependent data have been widely applied in classification problems in medicine (Hedeker, 2003) and numerous papers including Neuhas et al (1992), discuss methods for analysis of correlated binary and longitudinal data. This work provides feasibility to the analysis of random effects LR models for time dependent oil condition data. There is little evidence of other similar work in a condition-based maintenance or an engineering context.

3.3 Classification of data into four categories

Multinomial LR models have been developed to improve upon binary models by classifying oil samples into the four categories used at WesTrac. Table 1 is a classification table that highlights the number of correct and incorrect classifications when testing the model.

	Actual A	Actual B	Actual C	Actual X
Classified A	215	65	1	4
Classified B	83	487	85	5
Classified C	0	31	174	18
Classified X	0	2	10	40
Total	298	585	270	67
% Correct	72.2%	83.3%	64.4%	59.7%

Table 3: Classification Table of Multinomial LR Model

Of most significance is the small number of 'X sample' misclassifications. Of those classified as 'X', 0 of these are 'A samples', and only 2 are 'B samples'. Using these models to flag the more critical samples, it can allow those samples to be prioritised and assessed first, thus identifying machinery requiring immediate attention and action. This will allow WesTrac's oil laboratory to have a faster turnaround time on a large proportion of 'X samples', which will in turn contribute to the profitability of WesTrac client business.

5. Conclusions and Future Work

The results discussed demonstrate logistic regression models can not only be utilised to successfully classify oil samples, but also to determine and illustrate key relationships which exist between condition parameters and oil classification outcomes. Of most significance, the effect of previous compartment history on future classification is outlined. Additionally this work builds on previous work developing multinomial LR models to extend the classification to the four categories (A, B, C, X) currently in place at WesTrac.

Further work in relation to PCA provides promise to the development of models with higher classification accuracy than already present, thus allowing the models to play a key role in a move to a more automated classification process. Finally, the significance of the previous

compartment history on future performance provides feasibility for applying and developing random effects LR models accounting for the dependent nature of the oil data.

6. References

Aguilera, A.M., Escabias, M. & Valderrama, M.J. 2006. Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis,* 50**,** 1905-1924.

Hedeker, D., 2003. A mixed-effects multinomial logistic regression model. *Statistics in Medicine,* 22**,** 1433-1446.

Hosmer, D.W. & Lemeshow, S. 2000. *Applied Logistic Regression,* New York, John Wiley & Sons.

Janzen, F.J. & Stern, H.S. 1998. Logistic regression for empirical studies of multivariate selection. *Evolution,* 52**,** 1564-1571.

Kim, Y.S. 2010. Performance evaluation for classification methods: a comparative simulation study. *Expert systems with applications,* 37**,** 2292-2306.

Lin, D., Banjevic, D. & Jardine, A.K.S. 2006. Using principal components in a proportional hazards model with applications in condition-based maintenance, *The Journal of the Operational Research Society*, 57, 910-919.

Neuhaus, J. M. (1992). Statistical methods for longitudinal and clustered designs with binary data. *Statistical Methods in Medical Research*, 1, 249-273.

Ratnam, M. 2011. *Automatic "A Sample" Acknowledgement*. Mechanical Engineering, UWA.

Team, R. D. C. 2012. R: A Language and Environment for Statistical Computing. Vienna, Austria.

Williams, J.H., Davies, A. & Drake, P.R. 1995. Condition-based Maintenance and Machine Diagnostic, Chapman & Hall, London