Predictive analytics in incident prevention

Predictive analytics can support risk management by identifying where failures are likely to occur and what can be done to prevent them

WILLIAM R BROKAW
Kestrel Management

Companies are generating ever increasing amounts of data associated with business operations, leading to renewed interest in predictive analytics, a field that analyses large data sets to identify patterns, predict outcomes, and guide decision-making. Companies are also facing a complex and ever expanding array of operational risks to proactively identify and mitigate. While many companies have begun using predictive analytics to identify marketing/sales opportunities, similar strategies are less common in risk management, including safety.

Classification algorithms, one general class of predictive analytics, could be particularly beneficial to the refining and petrochemical industries by predicting the time frame and location of safety incidents based on safety related inspection and maintenance data, essentially leading indicators. There are two main challenges associated with this method: (1) ensuring that leading indicators being measured are actually predictive of incidents, and (2) measuring the leading indicators frequently enough to have predictive value.

A case study to illustrate this process is discussed in this article. Using regularly updated inspection data, the author developed a model to predict where broken rails are likely to occur in the railroad industry. The model was created using a logistic regression modified by Firth's penalised likelihood method, and predicts broken rail probabilities for each mile of track. Probabilities are updated as additional data are collected.

In addition to predicted broken rail probabilities, the model identifies the variables with the most predictive validity (those that significantly contribute to broken rails). Using the model results, the railroad was able to identify exactly where to focus maintenance, inspection, and capital improvement resources and what factors to address during these activities. Validation tests of the model revealed 70% of the actual broken rail incidents occurred on the 20% of segments at highest risk for broken rails.

The same methodology could be used in the refining and petrochemical industries to manage risks by predicting and preventing incidents, provided that organisations:

- Identify leading indicators with predictive validity
- Regularly measure leading indicators (inspection, maintenance, and equipment data)
- Create a predictive model based on measured indicators
- Update the model as data are gathered
- Use the outputs to prioritise maintenance, inspections, and capital improvement projects and review operational processes/practices.

Predictive analytics is a broad field encompassing aspects of various disciplines, including machine learning, artificial intelligence, statistics, and data mining. Predictive analytics uncovers patterns and trends in large data sets. One type of predictive analytics, classification algorithms, could be particularly beneficial to the refining and petrochemical industries.

Classification algorithms can be categorised as supervised machine learning. With supervised learning, the user has a set of data that includes predictive variable measurements which can be tied to known outcomes. In the model discussed in the case study section of this article, various track measurements (for instance, curvature, crossings) were taken during a two-year period for each mile of rail. The known outcome, in this case, is whether a broken rail occurred on each rail mile during that two-year period.

An appropriate modelling algorithm is then selected and used to analyse the data and to identify the relationships between the variable measurements and the outcomes to create predictive rules (a model). Once created, the model is given a new data set containing predictive variable measurements and unknown outcomes and will then calculate the outcome probability based on the model rules. This is in comparison to unsupervised learning types, in which algorithms detect patterns and trends in a data set with no specific direction from the user, other than the algorithm used.

Common classification algorithms include linear regression, logistic regression, decision tree, neural network, support vector machine/flexible discriminants, naïve Bayes classifier, and many more. Linear regressions provide a simple example of how a classification algorithm works. In a linear regression, a 'best-fit' line is calculated based on the existing data points, providing a $y = mx + b$ line equation. Inputting the known variable (x) gives a

prediction for the unknown variable (y).

Most real world relationships between variables are not linear, but complex and irregularly shaped. Therefore, linear regression is often not useful. Other classification algorithms are capable of modelling more complex relationships, such as curvilinear or logarithmic relationships. For example, a logistic regression algorithm can model complex relationships, can incorporate non-numerical variables (for instance, categories), and can often create realistic and statistically valid models. The typical output of a logistic regression model is predicted probabilities of the outcome/event occurring. Other classification algorithms provide a similar output as logistic regression, but the required inputs are different between algorithms.

The modelling of complex relationships is particularly useful in risk management, where risk is typically prioritised based on the likelihood and potential severity of a particular outcome. Modelling the risk factors that contribute to that outcome results in a precise and statistically valid estimate of outcome likelihood. In contrast, many risk assessments measure 'likelihood' on a categorical scale (once in ten years, once a year, multiple times per year), which is less precise, more subjective, and makes it impossible to distinguish between risks that are in the same broad category. Other techniques exist to quantifiably assess potential severity in a risk assessment, but that is beyond the purview of this article.

Case study

The author developed a predictive broken rail model for railroad application. Broken rails are a significant driver of derailment risk in railroad operations. Derailments caused by broken rails tend to have more severe consequences compared to other derailment causes since broken rail derailments typically occur at higher speed with little or no warning. The ability to predict where broken rails are likely to occur would allow for more effective management of broken rail derailment risk through targeted track inspections, maintenance, and capital improvement programmes.

The author developed and validated a predictive model of broken rail derailments on a mile-by-mile basis. The objectives were to:

- Identify the various factors that drive broken rail risk
- Quantify how each risk factor affects broken rail risk
- Develop a risk profile for each mile of track based on current and historical risk factors
- Translate the model results into easily understood language, thereby allowing field managers and engineers to prioritise corrective actions in real time based on current risk profiles.

Model development

The model development process followed these basic steps:

- Identification of potential predictor variables based on expert judgment
- Statistical evaluation of potential predictor variables for significance ($\alpha = 0.05$)
- Model development, evaluation, and selection
- Model output refinement
- Ongoing updates as new data are gathered.

Statistical model selection

Selection of an appropriate modelling methodology depends on the characteristics of the data that are being modelled. Review of the provided data indicated that:

- The predictor variables included both continuous (numerical) and categorical data (for example, class of rail, type of metal)
- The outcome variable was binary (that is, either a broken rail occurred or not)
- The number of null outcomes (no broken rail) significantly exceeded the number of broken rail events.

Because the outcome variable was binary, a logistic regression model was initially considered. However, the disparity between null outcomes and broken rail events presented a significant risk of biasing the model. As a result, a penalised likelihood logistic regression (using the Firth method) was chosen for modelling the broken rail data. The Firth method reduces bias resulting from disparities in outcomes, thereby allowing for a more representative model.

General model notes

All model development activities were performed in the statistical software program 'R' (v. 3.2.2) using the logistf package. Each variable was evaluated for validity using the p-value determined by R. The p-value gives the probability that any observed relationship between a predictor variable and the outcome variable is caused by random chance. Smaller p-values indicate that the relationship is not due to random chance, but rather to a valid predictive relationship between the two variables. Typically, p-values below 0.05 are considered statistically significant.

As with any statistical model, data accuracy is crucial for both the modelling process itself and for ongoing interpretation of the model. It is beneficial to review data collection processes to ensure that all model variables are measured and recorded accurately.

Variable identification

As the initial step in model development, subject matter experts within the railroad company were invited to brainstorm potential variables that influence whether a broken rail occurs. Emphasis was placed

Significant variables and p-values for broken rail model

Variable	p-value	Variable	p-value
Previous broken rails	~0.0	Track defects	1.503×10^{-8}
Track adjustments	~0.0	Curvature	4.038×10^{-7}
Welds	~0.0	Accumulated tonnage	~0.0
Rough rail (proprietary)	1.188×10^{-7}	Joints	2.496×10^{-10}

Table 1

on variables for which the relevant data were publicly available or gathered by the railroad company itself.

Modelling

A data set was created that associated each mile of track with measurements for each potentially predictive variable and the outcome variable (whether a broken rail occurred). The data set was modelled using the logistf package in 'R' (v. 3.2.2). Approximately 80% of the data set was used for training the model, while the other 20% was used to assess the accuracy of the trained model. The significant variables and corresponding p-values are given in **Table 1**.

Figure 1 shows a lift curve of the model predictions versus the performance of a random chance model. The greater the difference between the model performance and random chance performance, the more effectively the model is predicting broken rail occurrence. **Figure 1** suggests strong model performance compared to random chance.

Model outputs

The primary output of the modelling process consists of model coefficients for all significant predictive variables. These coefficients are combined into a model equation, which can be used to calculate the predicted probability of the model outcome (in this case, broken rails) based on the predictive variable measurements. Additional tools can be developed based on the coefficients and the predicted probabilities to aid in understanding the model results. For example, a tornado chart can be used to visually depict how each predictor variable in the overall model contributes to the outcome.

In the current Broken Rail Model, the following outputs were calculated for each mile of track in the system:

- Predicted broken rail probability
- Risk multiplier (how much more likely is a broken rail compared to the lowest risk mile?)
- Risk contribution for each predictor variable.

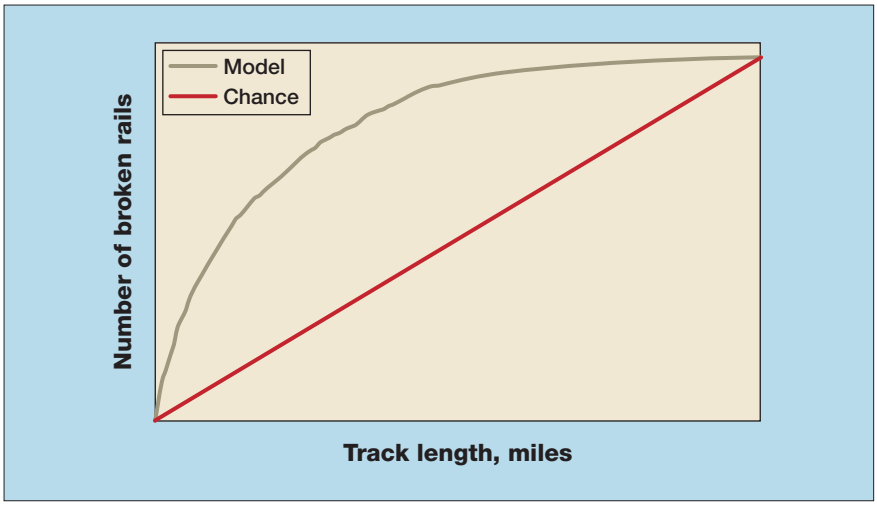

Figure 1 Lift curve for broken rail model

Predicted broken rail probabilities and risk multipliers are used to identify the track miles that have the highest risk of experiencing a broken rail. Risk contribution scores identify the predictor variables that must be addressed on each mile to reduce the overall broken rail risk. Using these scores, maintenance work and capital improvements can be prioritised to first address the most significant risk factors on the track miles with the highest overall risk.

Models could be created to predict the probability of failure for specific types of equipment within a refinery

Implications for petroleum refining and petrochemicals

While this case study focuses on application of predictive analytics to the railroad industry, the same general approach can be translated for use in petroleum related industries. Pipeline integrity programmes, for example, can benefit from predictive modelling approaches. Pipelines are similar to railroad tracks in that they are vital infrastructure spread over a large geographic area with significant consequences if failure occurs. Additionally, many companies are using specialised instruments to assess pipeline integrity. The data generated by these pipeline assessments provide a starting point for predictive modelling of pipeline failures.

While pipeline integrity is an obvious natural fit for this type of modelling, there are many potential EHS related applications as well, including mechanical integrity and human error risk. For example, models could be created to predict the probability of failure for specific types of equipment within a refinery, like pumps, valves, piping, boilers, vessels, or tanks.

The applicability of predictive modelling is limited by the data that are available. For example, the author has explored the application of predictive analytics for clients in the refining and petrochemical industries. The initial phase in each of these projects uncovered the importance of gathering a large amount of high quality quantitative data in a form appropriate for modelling. (Narrative text for what happened in a safety incident cannot be input into a model to yield predictive results.) Once adequate data are amassed, it becomes possible to turn them into a full and useful analytics programme.

Regardless of the application, the key considerations in assessing whether modelling may be appropriate include:

- What will the model predict (outcome of interest)?
- What factors increase or decrease the likelihood that the outcome of interest will occur (risk factors)?

• What data are available (internally or externally) that measure the outcome of interest and the risk factors?
• What is the quality of the data?
• What are the costs and benefits of gathering data that are not currently available?

There are many common pitfalls to completing an analytics project, but most involve insufficient consideration of the above. For example, many companies that are eager to begin an analytics project focus solely on using data they already have, without considering the quality, whether it is appropriate for the model they would like to build, or even if it is appropriate for modelling in general (for instance, free-form text). For this reason, advanced planning and a robust assessment of data quality and utility are all crucial to the success of an analytics project.

Conclusion

Much like the railroad case study, a classification algorithm could be used to predict the timeframe and location of safety related incidents based on leading indicators, provided that organisations:

• Identify risk factors with predictive validity (leading indicators)
• Regularly measure leading indicators
• Create a predictive model based on measured indicators
• Update the model as new indicator data are gathered
• Use the outputs to prioritise maintenance, inspections, and capital improvement projects and review operational processes/practices.

Companies in the refining and petrochemical industries are currently generating and recording unprecedented amounts of data associated with operations. Companies that strive to be best-in-class will need to use that data intelligently to guide business decision-making, particularly within the field of risk management. Predictive analytics can aid effective risk management by identifying where failures are likely to occur and what companies can do to prevent those failures.

The versatility of predictive analytics, including the method described in this case study, can be applied to help companies analyse a wide variety of problems. In this way, companies can explore and investigate past performance, gain the insights needed to turn vast amounts of data into relevant and actionable information, and create statistically valid models to facilitate data driven decisions.

William R Brokaw is a Consultant and Data Science Specialist with Kestrel Management. He is adept at using traditional statistical techniques, as well as predictive analytical methods, to draw conclusions from data that may not have been obvious through qualitative analysis. He has developed expertise on the comprehensive assessment of practices, processes, systems, and culture related to safety, and has a sound background in safety assessment that can be adapted and targeted to the needs of a particular client. He previously taught behavioural statistics and research methods at Norfolk State University.