MARY KAY O'CONNOR
PROCESS SAFETY CENTER
TEXAS A&M ENGINEERING EXPERIMENT STATION

19th Annual International Symposium
October 25-27, 2016 • College Station, Texas

Applying Predictive Analytics to Process Safety Leading Indicators

William R. Brokaw
Kestrel Management LLC

Author Email: wbrokaw@kestrelmanagement.com

Abstract

Leading indicators can be defined as safety-related variables that proactively measure organizational characteristics with the intention of predicting, and subsequently avoiding, process safety incidents. Leading indicators become especially powerful when combined with advanced statistical methods, including predictive analytics. Predictive analytics is a broad field encompassing aspects of various disciplines, including machine learning, artificial intelligence, statistics, and data mining.

This paper presents a case study in applying predictive analytics. **Methods:** The author developed a predictive derailment model for railroad application that could be modified and applied to process industries. Using regularly updated inspection data, the model was created using a logistic regression modified by Firth's penalized likelihood method due to the low ratio of events to misses (i.e. track miles with derailments compared to track miles without derailments). The resulting model provides derailment probabilities for each mile of track over a six-month period. Additionally, the model identifies the variables that are significantly contributing to derailments, thereby showing the company which factors to address to prevent future incidents. **Results:** Model validation revealed that it demonstrated statistically significant predictive ability for 75% of derailments. **Discussion:** The same methodology could be used in the process industries to predict and prevent incidents, provided that organizations:

1. Identify leading indicators with predictive validity
2. Measure indicators at regular intervals
3. Create a predictive model based on measured indicators
4. Deploy the model whenever leading indicator measurements are taken to calculate predicted incident probabilities

Introduction

In recent years, there has been significant research attention in developing and measuring leading indicators in the process safety industries. **Leading indicators** can be defined as safety-related variables that proactively measure organizational characteristics with the intention of predicting, and subsequently avoiding, process safety incidents. Conversely, **lagging indicators** are safety-related variables that have already occurred, such as total recordable incident rate (TRIR) or days away, restricted, or transferred duty rate (DART). Lagging indicators are helpful in documenting past patterns in safety performance; investigating their occurrence can help prevent future incidents through the implementation of lessons learned.

Implementing lessons learned from lagging indicators, however, requires an incident to occur first. Leading indicators are preferable because, if measured and analyzed correctly, they can help prevent incidents from ever occurring. Leading indicators become especially powerful when combined with the use of advanced statistical methods, including predictive analytics.

Predictive Analytics

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 branch of predictive analytics, classification algorithms, could be particularly beneficial to the process industries.

Classification algorithms can be categorized as supervised machine learning. With supervised learning, the user has a set of data that includes predictive variable measurements that can be tied to **known** outcomes. The algorithms identify the relationships between various factors and those outcomes to create predictive rules (i.e., a model). Once created, the model can be given a dataset with predictive variable measurements and **unknown** outcomes, and will then predict the outcome 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 modeling more complex relationships, such as curvilinear or logarithmic relationships. For example, a logistic regression algorithm can model complex relationships, can incorporate non-numerical variables (e.g. 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 to logistic regression, but the required inputs are different between algorithms.

Case Study

The author developed a predictive derailment model for railroad application. The railroad had experienced a number of derailments and, correspondingly, incurred costs over time. The author created a predictive model of track-caused derailments on a mile-by-mile basis to help predict and prevent future derailments. The objective was to apply statistically valid weighting factors and associate track measurements with derailment data to create a predictive model.

Penalized Likelihood Logistic Regression

Classification models learn predictive rules in an original data set that includes known outcomes, then apply the learned rules to a new data set to predict outcomes and probabilities. In this case study, a logistic regression modified by Firth's penalized likelihood method was used to:

- Fit the model;
- Identify the five significant predictive variables; and
- Calculate derailment probabilities for each mile of mainline track based on track measurements.

Data from a selected 18 month period were used to create and evaluate a pilot model. Subsequently, data from a 48 month period were used to create a refined model; finally, data from the following 12 months were used to evaluate the fit of the refined model.

Pilot Model

Using the data from the first 18 month period, a pilot model was created, which showed that the probability of a derailment happening on a particular mile of track by random chance is 0.08703%. Approximately 7% of the track miles had a predicted derailment probability above that 0.08703% threshold. 44% of the actual incidents occurred within the 7% of track miles with an elevated predicted probability of an incident.

To aid with prioritization, two risk matrices were created, both with predicted probability as the "likelihood" axis value. The hazard matrices were translated to numerical scores for each track mile by converting each variable to z-scores. Z-scores provide the relative position of each raw score in terms of the mean and standard deviation. This is primarily used for standardizing unlike distributions for the purpose of comparing them. Z-scores were calculated for the five significant predictive variables to aid with maintenance and capital improvement planning and resourcing.

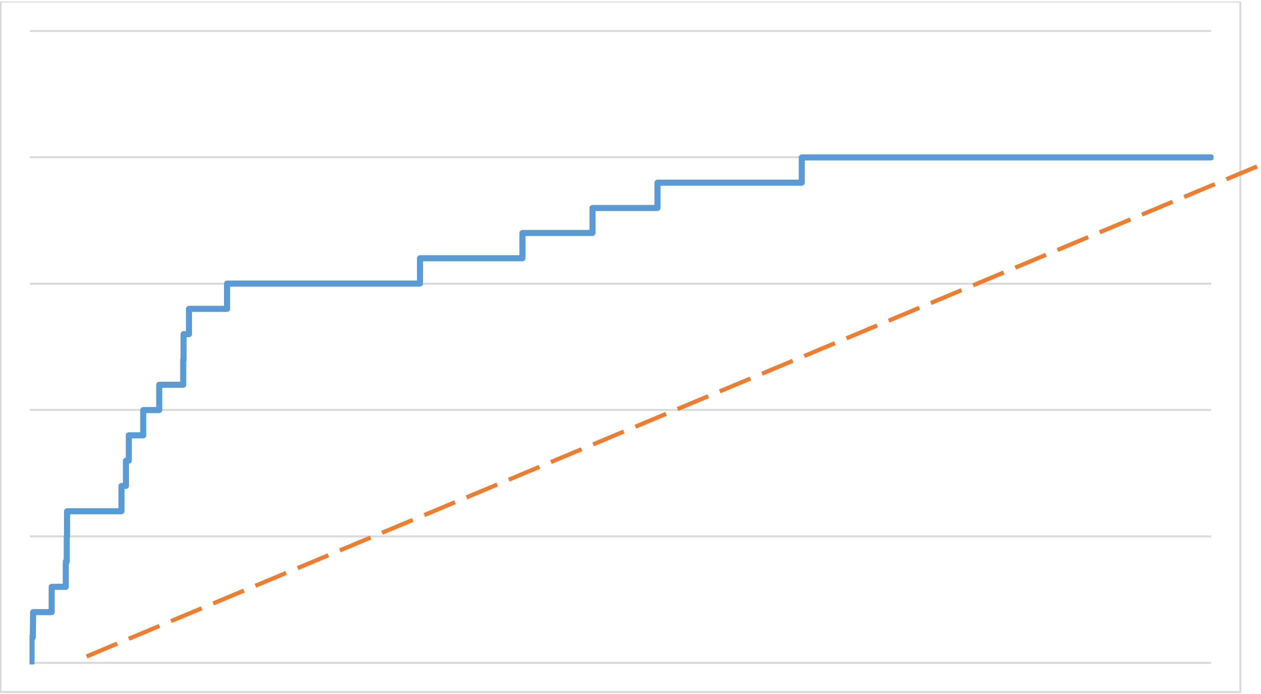
Refined Model

The model was subsequently refined using five years of data, as compared to the 18 months used in the pilot. The resulting model provides derailment probabilities for each mile of track over a six-month period.

The validation data set for the refined model was 12 months, so it was split into two six-month subsets for analysis:

- During the first six month validation: 19.8% of segments had elevated risk and 43.8% of incidents occurred on those segments.
- During the second six month validation: 20.6% of segments had elevated risk and 75% of incidents occurred on those segments. A receiver operating characteristic (ROC) curve for this evaluation period is included in Figure 2. A ROC curve shows the effectiveness of a model's predictions relative to random chance. The more space between the ROC curve and the line showing random chance, the more effective the model is at predicting events. *To maintain confidentiality, axis labels have been removed from Figure 2.*

Figure 2: 7/1/2015 Validation ROC Curve



Additionally, the model identified the five variables that significantly contribute to derailments, thereby showing the company which factors to address to help prevent future derailments.

Data Reconfiguration

The model outputs can be altered in various ways to make them more useful. For example, high-risk segments were sorted by subdivision to identify consecutive miles that are elevated. Average risk for each subdivision was calculated to determine highest risk areas for possibility of allocating larger capital improvement projects. In addition, significant predictive variables were examined individually to evaluate the need for corrective action.

Additional Tools

Since creating the model, a number of additional tools related to the predictive derailment model have been implemented to improve the usability of the data generated:

- An interactive Google Earth file shows the risk scores of every mile of track in the system.
- A layered Google Earth file identifies the top risks and the nearby mile posts that can be targeted during capital improvement projects.
- ROC curve shows that the model is a far better predictor of derailments than chance alone. It also shows that the greatest risk reduction for the investment may be obtained by focusing on the 2.5% of track miles with the highest probability of a derailment.
- Analysis of variance of each variable determines how changes to the variable scores affect predicted outcome probability.

- Averaged risk scores for each variable over ten-mile segments enables easier identification of opportunities for larger capital improvements.
- Averaged risk scores for each variable within a subdivision creates awareness of derailment risk and what variables to focus on when initiating prevention efforts.

Implications for the Process Industries

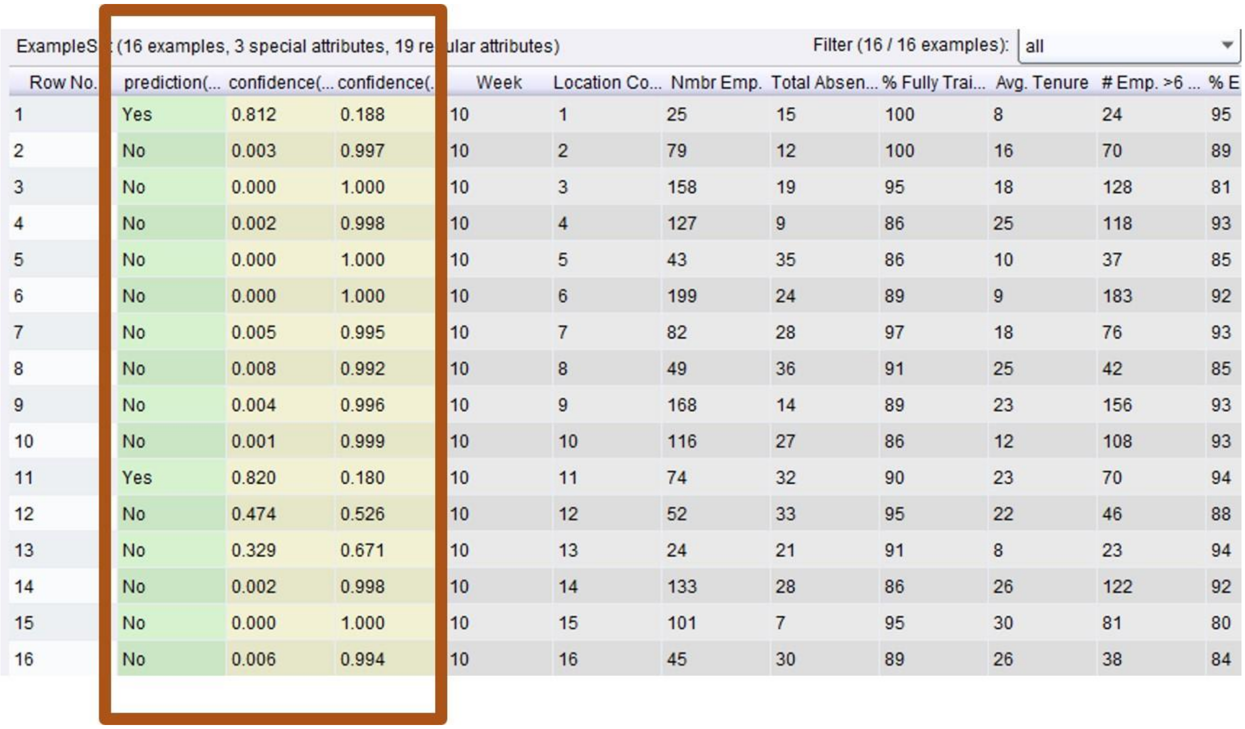
The same general approach described in the case study could be applied to the process industries. Leading safety indicators can be used as predictive variables and incidents could be used as the outcome. Figure 1 shows an example output from a logistic regression model applied to the process industries. The leading safety indicators (see columns 7-12 in Figure 1) are:

- Number of employees at the site,
- Total absences in the last week,
- Percentage of employees that are fully trained,
- Average employee tenure,
- Number of employees with more than six months of experience, and
- Number of employees with more than twelve months of experience.

Measurements for these variables would be taken regularly at each facility or unit. Precision increases as the measurements become more frequent and the observed area (facility/unit) becomes smaller. Once a sufficient amount of leading indicator measurements have been taken, they would then be combined with incident data to provide both the predictive variable measurements and the outcome data needed for training a model. This data set would be fed into a logistic regression or other classification algorithm to create a model.

Once the model has been created, it can be applied to new leading indicator measurements to predict the probability of an incident occurring at that location during the applicable time frame. Columns 2-6 of Figure 1 demonstrate this predictive capacity. Column 5 gives the time frame of the prediction, and column 6 states the applicable location number. Column 2 displays the predicted decision of the model (yes, an incident will occur; or no, an incident will not occur). Columns 3 and 4 show the confidence level associated with each potential decision. This essentially describes the predicted probability of the outcome happening or not happening. Once predicted incident probabilities have been found, management would be able to focus improvement resources on those locations that have the highest probabilities of experiencing an incident. The classification algorithms also identify which leading indicators have predictive validity, so management will know how improving those leading indicators will affect the predicted probability of incidents occurring. In other words, they will know which leading indicators have the strongest relationship will incidents, and can focus on improving those indicators first.

__Figure 1. Logistic Regression Predictive Model Example__



ExampleS (16 examples, 3 special attributes, 19 re ular attributes) Filter (16 / 16 examples): all

Row No.	prediction(...	confidence(...	confidence(.	Week	Location Co...	Nmbr Emp.	Total Absen...	% Fully Trai...	Avg. Tenure	# Emp. >6 ...	% E
1	Yes	0.812	0.188	10	1	25	15	100	8	24	95
2	No	0.003	0.997	10	2	79	12	100	16	70	89
3	No	0.000	1.000	10	3	158	19	95	18	128	81
4	No	0.002	0.998	10	4	127	9	86	25	118	93
5	No	0.000	1.000	10	5	43	35	86	10	37	85
6	No	0.000	1.000	10	6	199	24	89	9	183	92
7	No	0.005	0.995	10	7	82	28	97	18	76	93
8	No	0.008	0.992	10	8	49	36	91	25	42	85
9	No	0.004	0.996	10	9	168	14	89	23	156	93
10	No	0.001	0.999	10	10	116	27	86	12	108	93
11	Yes	0.820	0.180	10	11	74	32	90	23	70	94
12	No	0.474	0.526	10	12	52	33	95	22	46	88
13	No	0.329	0.671	10	13	24	21	91	8	23	94
14	No	0.002	0.998	10	14	133	28	86	26	122	92
15	No	0.000	1.000	10	15	101	7	95	30	81	80
16	No	0.006	0.994	10	16	45	30	89	26	38	84

Conclusion

Much like the railroad case study discussed above, a classification algorithm could be used to predict the timeframe and location of process incidents based on leading indicators. There are two main challenges associated with this strategy:

1. Ensuring that leading indicators are actually predictive of incidents, and
2. Measuring the leading indicators frequently enough for them to have predictive value.

Companies in the process 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 future business decision-making.

The versatility of predictive analytics, including the method described in this case study, can be applied to help companies analyze 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.