Executive Summary

This paper shows two methods of estimating the probability of an Gulf of Mexico OCS lease having an accident in the next year. These methods use public data. The first method is a categorical approach. It uses 6 yes or no attributes of the lease. Such as was there production from 4 or more completions during the year. For the leases with 5 yes attributes, it is over 30 times more likely to have an accident than a lease with 0 yes attributes. The second method is more complex utilizing a specialized regression to estimate lease probability. This method employs 8 predictors of future accidents. An example of the predictors is the count of the number of accidents.

Data mining was used in development of these methods. Data mining is the fusion of multivariate statistics, database management systems, and artificial intelligence to find patterns in data. About 50 potential predictors were considered. The data was gathered and arranged so it could be analyzed. The data then was randomly split into 2 parts. The first part was a training set. The training set was about 70% of the total. All the various data mining approaches were applied to the training set. They yielded the two methods. Then those 2 methods were applied to the remaining 30% of data (test data). To see if the methods were actually effective on independent data. Both methods obtained similar levels of effectiveness in estimating the probability of an accident next year.

Bureau of Safety and Environmental Enforcement (BSEE) has a stated goal of risked based decision making. These methods compute those risks. BSEE holds more data than was used in this analysis. It is reasonable that the complete BSEE data would yield even better methods with additional data mining.

OCS Accident Forecasting Using Data Mining

April 2014

Introduction

The *Deepwater Horizon* accident increased the interest in accident prevention on the Outer Continental Self (OCS). One idea suggested by the Department of the Interior's Ocean Energy Safety Advisory Committee for the Bureau of Safety and Environmental Enforcement (BSEE) was an early warning system for accidents. Such systems would fit within the BSEE goal of risked based decision making. This paper will present two algorithms for forecasting future OCS accidents. These algorithms were developed using data mining procedures. Data mining has been successful in a number of settings. Further the effectiveness of these algorithms will be validated.

Data

All the raw data used in this paper was downloaded from the BSEE website, except the Incident of Non-Compliance (INCS) data which was obtained via Freedom of Information Request (BSEE-2013-00182). The INCS data is posted at this address (http://tedspublicpolicy.weebly.com/research-projects.html).

Six aggregations of data were used in this study:

Accident History
Sale Bid History
INCS Data
Lease Attributes
Drilling History
Production History

The historical span of this study is bounded in two ways. First by the period 1996 to 2013. Second for the specific lease from the year of first well spud until the year of lease expiration. All the data is aggregated into year and lease (YR_Lease) records for analysis. This analysis is limited to OCS activities in the Gulf of Mexico.

The approach is to use the attributes for a specific YR_Lease to project the relative likelihood this lease will have an accident in the next year. Such as using data associated with 1997 for G12345 and estimating the potential of an accident in 1998 for G12345. That implies projections for next years accidents will be created for 1997 through 2013. Two kinds of annual forecasts of accidents were developed.

In building an algorithm there are two competing trade offs. First is to add more variables to obtain additional precision. There is a potential for overfitting. That is the algorithm is capturing the random noise as opposed to the underlying relationship. The second trade off is the desire for parsimony. That the algorithm be simple.

The data set is spilt into two parts. A training set which has about 70% of the leases. The training data will be used to develop the algorithms. Upon completion of the analysis. The algorithms are tested against the remaining 30% of the data to validate the algorithm. Using this process design overfitting would be exposed, if it exists.

Categorical Algorithm

The Categorical Algorithm is a classification process. It partitions the YR_Lease combinations into categories ranging from low risk to high risks. For the training set the probability of a YR_Lease experiences an accident in the next year is approximately 2%. The partitions are made by using a series of yes and no questions. Appendix 1 contains the lists of all the potential indicators of future accidents. There were nearly 50 indicators considered. The number of yes attributes for a YR_Lease determines the partition placement. The data mining analysis yielded a set of six questions:

1. Was there an INC during the year? (INCS_Flag)
2. Was there an INC Number of type E100, G110, P103, or P451 during the year? (INCS_List)
3. Was there an accident during the year? (Accident_Flag)
4. Was there production from 4 or more completions during the year? (4Completion_Flag)
5. Was there 3 or more wells spud in the year? (3Well_Flag)
6. Was the lease located in any of these Area Codes (WR, KC, PL, GC, MC, SP, SS, MP, EW)? (AC_Flag)

Note that the questions come from 5 of the 6 aggregations of data. Only the Sale Bid History was not used as a source and the INCS Data provided two of the questions.

Categorical Algorithm applied to the training set

Number Yes	Leases with Accident Next Year	Leases without Accident Next Year	Percentage
0	93	11093	0.83
1	161	9177	1.72
2	128	3487	3.54
3	87	1295	6.30
4	42	391	9.70
5	24	66	26.67
6	1	9	10.00

The categorical algorithm is impressive. For YR_Leases with 0 yes response to the 6 questions the percentage of YR_Leases with an accident next year is less than 1%. As the number of yes responses increases so does probability of an accident. 6 yes is rare. It does break the pattern increasing of probability. Even at 10% it is the second highest percentage. A YR_Lease with 5 yes is over 32 times more likely to have an accident next year than a YR_Lease with no yes responses.

Logistic Regression Algorithm

A standard statistic approach for estimation is a linear regression. A linear regression can have a form like this:

$$Y = aX + bW + cZ + k + error$$

where Y is the attribute being estimated. X, W, and Z are independent variables. The parameters a, b, c, and k are found via minimizing least square errors. In the context here Y is a probability, such as the probability a YR_Lease will have an accident next year. Linear regression can not be applied because probabilities are bounded by zero and one. In statistics there is a method known as Logistics Regression that permits the estimating a probabilities. It requires some mathematical gymnastics.

Step 1: P = Probability of an event (such as an accident next year for a lease)

Step 2: Compute the odds of the event = P / (1 - P). Odds is a real positive number.

Step 3: Take the log of the odds = log(P / (1 - P)). Log of odds is a real number without bounds. This quantity can be used in a linear regression.

Namely:

$$\log(P / (1 - P)) = aX + bW + cZ + k + error$$

Again the indicator variables of Appendix 1 were consider for the algorithm. That process yielded this equation:

$$\begin{aligned}\log(P / (1 - P)) = \quad & 0.60198 * Accident\_Count + 0.33082 * 10Min\_Bid \\ & + 0.10158 * Well\_Count + 1.00322 * INCS\_Flag \\ & + 0.50159 * E100 + 0.66806 * AC\_Flag \\ & + 0.47252 * Oil\_Flag + 0.67412 * 4Completion\_Flag \\ & - 5.16431\end{aligned}$$

This approach generates an estimate of probability of an accident next year for a specific YR_Lease. To display the effectiveness of the Logistic Regression the probability percentages estimated are rounded and placed into a Probability Percentage Interval for tabulation of results. As an example the Logistic Regression probability percentage estimate of 2.8% would be placed into the 3% Probability Percentage Interval.

The following table shows the fit of the Logistic Regression:

Logistic Regression Probability Percentage Interval	Actual Probability Percentage
1	0.84
2	1.94
3	2.54
4	4.28
5	4.96
6	6.09
7	9.70
8	8.97
9	10.38
10	8.97
11	15.80
12	13.27
13	7.69
14	20.00
15	18.52
16	25.00
17	25.00
18	10.53
19	25.00
20	0.00

Application of Algorithms to Test Data Set

About 70% of leases were used to develop the two algorithms. Now those algorithms can be applied to the other 30% to test the effectiveness of both algorithms.

The comparison for the Categorical Algorithm:

Number of Yes	Training Data Percentage	Test Data Percentage
0	0.83	0.77
1	1.72	1.83
2	3.54	3.98
3	6.30	5.05
4	9.70	12.82
5	26.67	27.78
6	10.00	16.67

The Categorical Algorithm performs in a very similar pattern in both the training and independent test data set. In this case the ratio of probabilities for 5 yes compared to 0 yes is over 36.

The comparison for the Logistic Regression:

Logistic Regression Probability Percentage Interval	Training Data Actual Probability Percentage	Test Data Actual Probability Percentage
1	0.84	0.82
2	1.94	2.13
3	2.54	1.95
4	4.28	6.16
5	4.96	6.22
6	6.09	6.47
7	9.70	8.25
8	8.97	11.11
9	10.38	9.09
10	8.97	7.69
11	15.80	13.33
12	13.27	5.88
13	7.69	18.75
14	20.00	5.00
15	18.52	15.38
16	25.00	0.00
17	25.00	28.57
18	10.53	0.00
19	25.00	50.00
20	0.00	20.00

Again the algorithms perform in a similar manner for the training and test data sets. At the higher probability percentage intervals the number of YR_Lease observations is smaller which leads to a larger spread in the actual probabilities.

Deepwater Horizon

If available data from 2009 were applied to G32306 in 2010 (*Deepwater Horizon*). Then the following estimates are generated:

For the Categorical Algorithm there are 4 yes. The training set had a probability of an accident next year for a YR_Lease with 4 yes of 9.7% For the Logistic Regression Algorithm, estimated accident probability of 12.1%.

Summary

The goal of this research was to determine if data mining technology could estimate the probability of a lease will have an accident next year based solely on public data from the prior year. Two algorithms have been presented that meet this goal. The categorical algorithm creates a partitions of leases where the probability of an accident can differ by a factor of over 30.

BSEE holds more data than was used here. It is reasonable to expect with that additional data, better algorithms would be developed. BSEE asserts it is taking a risk based decision making. These algorithms computes those risks.

Appendix 1

Variables used in the study

Database Attribute Name	Definition
YR_Lease	Catenation of Year and LEASE_NUM (Primary Key for Study)
YR	Year
LEASE_NUM	Lease Number
Next_YL	YR_Lease for next year for this LEASE_NUM
Last_Digit	Last Digit of LEASE_NUM (used to partition data into training & test sets; 0,3,5 are for test set)
Accident_Flag	1 if at least one accident during year
BID_AMOUNT	Bonus Bid Amount at Sale
Platforms	Number of Installed Platforms
FIRST_PRODUCTION_DATE	First Production Date
AREA_CODE	Area Code
End_YR	Year of lease expiration, if none set to 9999
1st_SPUD_YR	Year of first well spud
Accident_Count	Number of accidents in year
Panel_Count	Number of Panel accidents in a year
Mil_Bid	1 if BID_AMOUNT > 1,000,000
10Mil_Bid	1 if BID_AMOUNT > 10,000,000
Num_Bids	Number of bids in sale
Age	Years since 1st Spud (1st_SPUD - YR)
Well_Count	Number of Wells in year
OIL	Oil Production in BBLs in year
GAS	Gas Production in MCF in year
BOE	Barrel Oil Equalivant production in BBLs in year
Top_OPER_NUM	OPERATOR_NUMBER for top operators, 00000 for others
Completions	Number of producing completions
INCS_Count	Count of INCS
INCS_Flag	1 if INCS_Count > 0

E100	1 if there is an E100 INC
G110	1 if there is an G110 INC
G111	1 if there is an G111 INC
G112	1 if there is an G112 INC
G115	1 if there is an G115 INC
G231	1 if there is an G231 INC
P103	1 if there is an P103 INC
P240	1 if there is an P240 INC
P241	1 if there is an P241 INC
P280	1 if there is an P280 INC
P283	1 if there is an P283 INC
P401	1 if there is an P401 INC
P404	1 if there is an P404 INC
P406	1 if there is an P406 INC
P411	1 if there is an P411 INC
P412	1 if there is an P412 INC
P422	1 if there is an P422 INC
P431	1 if there is an P431 INC
P451	1 if there is an P451 INC
P470	1 if there is an P470 INC
Age24	1 if Age > 23
AC_Flag	1 if AREA_CODE in set {WR, KC, PL, GC, MC, SP, SS, MP, EW}
Oil_Flag	1 if OIL >= 0.5*BOE
BOE_Flag	1 if BOE > 100,000
Well_Flag	1 if Well_Count > 0
3Well_Flag	1 if Well_Count > 2
4Completion_Flag	1 if Completion_Count > 3
OPER_Flag	1 if Top_OPER_NUM is in set { 00078, 00105, 00698, 00724, 00540 }
INC Number	A number assigned to a violation of federal regulations
INCS_List	1 if INC Number is in {E100, G110, P103, P451}