STAT 444/CM 464 Final Project

Curtis Bright

Insurance fraud detection is a problem of obvious importance, and may potentially be aided through database analysis by an automated computer program. In this project, data from 6142 automobile owners in 1994 was used to identify the fradulent "risk" of 9278 automobile owners in 1995 and 1996.

The given data consisted of 33 variables, which included *Policy number* and *Year* (not used for prediction purposes) as well as *Fraud found*, a variable which indicated known fradulent activity (the response variable). The remaining 30 variables could be used to determine "high-risk" from "low-risk" individuals for 1995 and 1996, i.e., they could provide an estimatation of *Fraud found* based on the known 1994 values.

The key assumption used in calculation of this estimate is that *past tendencies are likely to be correlated with future tendencies*. Thus, a simple method to estimate *Fraud found* in January 1995 would be to look at the tendency of fraudulent activity to occur in January 1994: the hope is that those causes which influence *Fraud found* in January are likely to repeat in January of the following years.

Examining the 1994 data shows that the month of the year indeed has some correlation with fradulent activity: out of 470 accidents in August, 56 of them were fraudulent (11.9%). However, out of 447 accidents in October, only 1 was fradulent (0.2%).

It seems reasonable to use this statistic to help predict future fradulent claims. In fact, each of the 30 predictor variables can be examined in this way; the complete statistics are given at the end of this report, generated by the SQL queries like the following:

```
SELECT A.Month, A.Total, A.Total-B.NoFraud AS Frauds
FROM
  (SELECT Month, Count(*) AS Total
  FROM Learning
  GROUP BY Month) AS A
 JOIN
  (SELECT Month, Count(*) AS NoFraud
  FROM Learning
  WHERE FraudFound=0
  GROUP BY Month, FraudFound) AS B
 ON A.Month=B.Month;
```

These statistics can then be used to assess an individual's fradulent "risk": someone who has many attributes which historically have a high proportion of fraudulent activity can be reasoned to have a high fraud risk. Since the evaluation Gain function `avgp` is invariant to monotonic transformations of the probability estimate, we do not have to scale the the risk predictions to be between 0 and 1. Instead, we may simply take the sum of the historic porportions for each group the individual belongs to.

This method has the advantage that it is simple to understand, that the every predictor variable contributes to the risk assesment, and that a "high-level" understanding of the variables is not required. However, the method works best with *ordinal* variables, not continuous ones—it might not be expected to work well with a variable like *Age*. For example, 11.1% of claims from 74 years olds were fraudulent, but 0% of those from 75 year olds. Should there really be that much of a difference between the two?

Therefore, it is possible the fradulent proportion of some predictor variables is more "random" than useful. To test this, the 1994 data was split into two random subsets and the the fraud proportions from one set was used to predict the fraud risk of the individuals of the other set. The accuracy was evaluated using the Gain function and then compared to the accuracy when one of the 30 variables was removed from data. In fact, with the *Age* variable removed performance improved from 0.205 to 0.230 in one case and from 0.161 to 0.175 in the other.

Finally, using the variables determined to be "useful", it was possible to estimate a fraud risk for the 1995 and 1996 data using the known fraudulent proportions from 1994.

Month	Total	Frauds
Apr	533	47 (8.8%)
Aug	470	56 (11.9%)
Dec	471	17 (3.6%)
Feb	528	36 (6.8%)
Jan	608	48 (7.9%)
Jul	495	32 (6.5%)
Jun	543	47 (8.7%)
Mar	584	56 (9.6%)
May	569	52 (9.1%)
Nov	453	8 (1.8%)
Oct	447	1 (0.2%)
Sep	441	9 (2.0%)

WeekOfMonth	Total	Frauds
1.0	1281	100 (7.8%)
2.0	1387	82 (5.9%)
3.0	1448	90 (6.2%)
4.0	1369	94 (6.9%)
5.0	657	43 (6.5%)

DayOfWeek	Total	Frauds
Friday	999	69 (6.9%)
Monday	1053	69 (6.6%)
Saturday	761	54 (7.1%)
Sunday	688	59 (8.6%)
Thursday	865	49 (5.7%)
Tuesday	933	57 (6.1%)
Wednesday	843	52 (6.2%)

Make	Total	Frauds
Accura	202	29 (14.4%)
BMW	4	0 (0.0%)
Chevrolet	679	45 (6.6%)
Dodge	41	1 (2.4%)
Ford	182	17 (9.3%)
Honda	1147	79 (6.9%)
Jaguar	4	0 (0.0%)
Mazda	935	50 (5.3%)
Mecedes	2	0 (0.0%)
Mercury	36	3 (8.3%)
Nisson	12	1 (8.3%)
Pontiac	1489	90 (6.0%)
Porche	2	0 (0.0%)
Saab	48	4 (8.3%)
Saturn	22	3 (13.6%)
Toyota	1232	84 (6.8%)
VW	105	3 (2.9%)

AccidentArea	Total	Frauds
Rural	642	73 (11.4%)
Urban	5500	336 (6.1%)

DayOfWeekClaimed	Total	Frauds
0	1	0 (0.0%)
Friday	1019	81 (7.9%)
Monday	1496	97 (6.5%)
Saturday	53	0 (0.0%)
Sunday	21	0 (0.0%)
Thursday	1035	57 (5.5%)
Tuesday	1366	95 (7.0%)
Wednesday	1151	79 (6.9%)

MonthClaimed	Total	Frauds
0	1	0 (0.0%)
Apr	528	52 (9.8%)
Aug	467	57 (12.2%)
Dec	395	5 (1.3%)
Feb	534	38 (7.1%)
Jan	607	46 (7.6%)
Jul	488	25 (5.1%)
Jun	529	51 (9.6%)
Mar	596	45 (7.6%)
May	596	58 (9.7%)
Nov	482	6 (1.2%)
Oct	475	8 (1.7%)
Sep	444	18 (4.1%)

WeekOfMonthClaimed	Total	Frauds
1.0	1390	107 (7.7%)
2.0	1491	99 (6.6%)
3.0	1439	96 (6.7%)
4.0	1345	79 (5.9%)
5.0	477	28 (5.9%)

Sex	Total	Frauds
Female	954	40 (4.2%)
Male	5188	369 (7.1%)

MaritalStatus	Total	Frauds
Divorced	30	2 (6.7%)
Married	4189	279 (6.7%)
Single	1911	128 (6.7%)
Widow	12	0 (0.0%)

Fault	Total	Frauds
Policy Holder	4508	392 (8.7%)
Third Party	1634	17 (1.0%)

PolicyType	Total	Frauds
Sedan - All Perils	1664	216 (13.0%)
Sedan - Collision	2209	143 (6.5%)
Sedan - Liability	1922	15 (0.8%)
Sport - All Perils	9	0 (0.0%)
Sport - Collision	174	19 (10.9%)
Sport - Liability	1	0 (0.0%)
Utility - All Perils	142	15 (10.6%)
Utility - Collision	10	1 (10.0%)
Utility - Liability	11	0 (0.0%)

VehicleCategory	Total	Frauds
Sedan	3873	359 (9.3%)
Sport	2106	34 (1.6%)
Utility	163	16 (9.8%)

VehiclePrice	Total	Frauds
20000 to 29000	3192	169 (5.3%)
30000 to 39000	1387	86 (6.2%)
40000 to 59000	165	14 (8.5%)
60000 to 69000	31	1 (3.2%)
less than 20000	400	44 (11.0%)
more than 69000	967	95 (9.8%)

RepNumber	Total	Frauds
1.0	383	29 (7.6%)
2.0	401	25 (6.2%)
3.0	397	21 (5.3%)
4.0	370	31 (8.4%)
5.0	415	22 (5.3%)
6.0	368	32 (8.7%)
7.0	406	32 (7.9%)
8.0	399	24 (6.0%)
9.0	372	28 (7.5%)
10.0	388	32 (8.2%)
11.0	379	17 (4.5%)
12.0	372	22 (5.9%)
13.0	331	25 (7.6%)
14.0	392	23 (5.9%)
15.0	381	18 (4.7%)
16.0	388	28 (7.2%)

Deductible	Total	Frauds
300.0	3	0 (0.0%)
400.0	5911	379 (6.4%)
500.0	105	21 (20.0%)
700.0	123	9 (7.3%)

DriverRating	Total	Frauds
1.0	1531	109 (7.1%)
2.0	1530	97 (6.3%)
3.0	1558	97 (6.2%)
4.0	1523	106 (7.0%)

Days:Policy-Accident	Total	Frauds
1 to 7	8	1 (12.5%)
15 to 30	23	2 (8.7%)
8 to 15	14	1 (7.1%)
more than 30	6077	402 (6.6%)
none	20	3 (15.0%)

Days:Policy-Claim	Total	Frauds
15 to 30	23	2 (8.7%)
8 to 15	8	2 (25.0%)
more than 30	6110	405 (6.6%)
none	1	0 (0.0%)

PastNumberOfClaims	Total	Frauds
1	1464	109 (7.4%)
2 to 4	2170	132 (6.1%)
more than 4	765	18 (2.4%)
none	1743	150 (8.6%)

AgeOfVehicle	Total	Frauds
2 years	25	1 (4.0%)
3 years	65	6 (9.2%)
4 years	100	7 (7.0%)
5 years	558	44 (7.9%)
6 years	1356	85 (6.3%)
7 years	2312	152 (6.6%)
more than 7	1550	84 (5.4%)
new	176	30 (17.0%)

AgeOfPolicyHolder	Total	Frauds
16 to 17	155	29 (18.7%)
18 to 20	11	0 (0.0%)
21 to 25	50	8 (16.0%)
26 to 30	256	13 (5.1%)
31 to 35	2210	153 (6.9%)
36 to 40	1650	108 (6.5%)
41 to 50	1090	59 (5.4%)
51 to 65	537	25 (4.7%)
over 65	183	14 (7.7%)

PoliceReportFiled	Total	Frauds
No	6000	405 (6.8%)
Yes	142	4 (2.8%)

WitnessPresent	Total	Frauds
No	6103	406 (6.7%)
Yes	39	3 (7.7%)

AgentType	Total	Frauds
External	6059	409 (6.8%)
Internal	83	0 (0.0%)

NumberOfSuppliments	Total	Frauds
1 to 2	967	57 (5.9%)
3 to 5	825	39 (4.7%)
more than 5	1481	80 (5.4%)
none	2869	233 (8.1%)

AddressChange-Claim	Total	Frauds
1 year	68	7 (10.3%)
2 to 3 years	116	24 (20.7%)
4 to 8 years	251	17 (6.8%)
no change	5703	358 (6.3%)
under 6 months	4	3 (75.0%)

NumberOfCars	Total	Frauds
1 vehicle	5698	372 (6.5%)
2 vehicles	283	23 (8.1%)
3 to 4	149	14 (9.4%)
5 to 8	10	0 (0.0%)
more than 8	2	0 (0.0%)

BasePolicy	Total	Frauds
All Perils	1815	231 (12.7%)
Collision	2393	163 (6.8%)
Liability	1934	15 (0.8%)

Age	Total	Frauds
0.0	155	29 (18.7%)
16.0	7	0 (0.0%)
17.0	4	0 (0.0%)
18.0	22	2 (9.1%)
19.0	13	3 (23.1%)
20.0	15	3 (20.0%)
21.0	66	2 (3.0%)
22.0	47	6 (12.8%)
23.0	51	0 (0.0%)
24.0	48	2 (4.2%)
25.0	44	3 (6.8%)
26.0	198	14 (7.1%)
27.0	211	11 (5.2%)
28.0	231	12 (5.2%)
29.0	209	11 (5.3%)
30.0	241	15 (6.2%)
31.0	232	15 (6.5%)
32.0	210	24 (11.4%)
33.0	208	17 (8.2%)
34.0	232	14 (6.0%)
35.0	238	20 (8.4%)
36.0	158	10 (6.3%)
37.0	166	12 (7.2%)
38.0	161	9 (5.6%)
39.0	173	12 (6.9%)
40.0	152	18 (11.8%)
41.0	174	11 (6.3%)
42.0	168	7 (4.2%)
43.0	163	13 (8.0%)
44.0	170	8 (4.7%)
45.0	165	8 (4.8%)
46.0	98	7 (7.1%)
47.0	132	3 (2.3%)
48.0	119	6 (5.0%)
49.0	108	2 (1.9%)
50.0	125	9 (7.2%)
51.0	124	9 (7.3%)
52.0	95	6 (6.3%)
53.0	87	3 (3.4%)
54.0	91	5 (5.5%)
55.0	111	9 (8.1%)
56.0	52	1 (1.9%)
57.0	55	4 (7.3%)
58.0	48	3 (6.3%)
59.0	50	0 (0.0%)
60.0	70	2 (2.9%)
61.0	61	5 (8.2%)
62.0	41	2 (4.9%)
63.0	51	2 (3.9%)
64.0	56	2 (3.6%)
65.0	53	4 (7.5%)
66.0	18	3 (16.7%)
67.0	15	2 (13.3%)
68.0	10	1 (10.0%)
69.0	12	0 (0.0%)
70.0	8	0 (0.0%)
71.0	17	0 (0.0%)
72.0	16	2 (12.5%)
73.0	12	1 (8.3%)
74.0	9	1 (11.1%)
75.0	7	0 (0.0%)
76.0	18	1 (5.6%)
77.0	10	0 (0.0%)
78.0	10	1 (10.0%)
79.0	6	1 (16.7%)
80.0	15	1 (6.7%)