

POINT OF SALE FRAUD PREVENTION

Student team: Jeff Grossman, Dawn Herndon, Andrew Kaplan, Mark Michalski, Carlton Pate

Faculty Advisors: Peter Beling and James Lark
Department of Systems Engineering

Client Advisor: The Auto-Insurance Company

KEYWORDS: Fraud, Logistic Regression, Point of Sale (POS), Special Investigative Unit (SIU).

ABSTRACT

Fraud is an estimated \$20 billion annual expense to the auto-insurance industry, and no one is attempting to predict it at the point of sale (POS). Therefore, the Auto-Insurance Company and the Capstone team completed this project to combat fraud by detecting POS data characteristics that indicate fraudulent behavior. The objective of this project was to identify data characteristics of potential customers at the POS that correlate with fraudulent behavior.

A logistic regression model, created in S-plus, shows the significance of numerous POS data fields for fraud prediction. The model calculates the probability that a potential customer will be fraudulent. The Capstone team analyzed the model output to determine its accuracy and an optimal threshold probability. The logistic model created for the Auto-Insurance Company evaluates insurance applicants at the POS and determines the amount of risk they present to the company.

The proposed model has the potential to impact the Auto-Insurance Company, other insurance companies, and consumers on a variety of levels. Implementation of the model could minimize the Auto-Insurance Company's losses suffered due to fraud, thus providing incentives to lower rates charged to consumers. The Capstone team encourages the Auto-Insurance Company to continue to research and begin to implement predictive technology to meet the constantly changing demands of fraud.

INTRODUCTION

Fraud is one of the biggest problems facing auto-insurance companies today. In the U.S., auto-insurance companies incur costs of \$20 billion per year due to fraud. This means that the average annual auto-insurance premium is \$200 to \$300 higher because of

fraud (Wright 2001). The Auto-Insurance Company is an automobile-insurance provider who strives to give quick and fair premium quotes via the Internet, by telephone, or through independent agents.

Obviously, the goal of the Auto-Insurance Company is to make a profit while providing fair rates to customers. After achieving an underwriting profit over the five years prior to 2000, the Auto-Insurance Company claimed an underwriting loss in 2000. Fraudulent claims account for some of the underwriting losses every year. Therefore, it would be beneficial for the Auto-Insurance Company to reduce the number of fraudulent claims filed.

There is no current analysis to identify potentially fraudulent customers at the Auto-Insurance Company. The purpose of this project was to combat fraudulent property damage, specifically organized theft and arson, at the POS. The Capstone team created a statistical model that uses POS data to calculate the probability that a potential customer will file a fraudulent theft or fire.

PROBLEM STATEMENT

Current fraud investigation techniques at the Auto-Insurance Company occur after a claim has been filed, once the fraud has already occurred. The fraud investigation process at the Auto-Insurance Company involves a claims agent gathering data from the claimant while a response vehicle goes to the accident scene or wherever is convenient for the claimant. If a series of flags or indicators are identified at the accident scene or through the questioning process, a Special Investigative Unit (SIU) responds to the accident scene for further investigation. The SIU determines if the claim is fraudulent or truthful. SIUs have expert knowledge and prove to be extremely useful, but are an expensive step in the Auto-Insurance Company's claims process.

Claims are paid out accurately and quickly because that is the nature of the insurance industry, and the Auto-Insurance Company's reputation depends on

efficient processes. Therefore, an SIU must be certain that a claim is fraudulent before filing a lawsuit or not paying the claim. The Auto-Insurance Company has much more to lose if they assume a truly accurate claim to be fraudulent than if they pay out a fraudulent claim. Also, it raises the possibility that a claimant can sue the insurance company claiming bad faith, which will allow the claimant to pursue punitive damages.

The goal of the Capstone team was to provide the Auto-Insurance Company with a model that assesses a potential customer's risk for fraud at the POS. The model identifies which customers are most likely to commit organized theft or arson and gives the Auto-Insurance Company the opportunity to charge them a higher premium or deny them coverage altogether. With this model, the costs of fraud can be drastically reduced or avoided through early identification.

METHODOLOGY

A model that uses POS information to calculate the probability of a customer committing fraud is a new idea in the auto-insurance industry. It required extensive data manipulation, fraud investigation, and logistic regression.

The data for this project was originally in multiple files containing about 20,000 policies and 350 fields. The policy number linked all the files.

To begin the project, the Capstone team had to understand what data the Auto-Insurance Company collects at the POS. POS data is stored in three files: policy, driver, and vehicle. The policy file contains data specific to each policy, the driver file contains data about each driver on each policy, and the vehicle file contains data about each vehicle on each policy. The majority of the data in these files is available when a customer buys auto insurance.

Data Manipulation

Manipulating the data to a format that the Capstone team could run logistic regressions was the first step of this project. The Capstone team divided data manipulation into three phases.

- Combine the policy, driver, and vehicle data into one file using Microsoft Access. This step made each observation a unique combination of policy, driver, and vehicle data. The resulting sample size was greater than 65,000.
- Eliminate the redundant fields, fields that contain incomplete data, and fields that are not available at the POS from this file. The resulting number of fields shrunk to about 100.

- Code all the categorical variables into formats suitable for logistic regression.

Fraud Investigation

The Capstone team could not create a model to predict fraud without information about existing fraudulent claims. Unfortunately, the Auto-Insurance Company does not collect data indicating which claims were fraudulent. Thus, the Capstone team had a sample of 200 claims reinvestigated by an SIU. The SIU returned the claims indicating which claims they thought were likely fraudulent and which claims they thought were not fraudulent. The Capstone team created a new "likely fraud" field in its data file, which indicated an observation that was likely to be fraudulent with a 1 and an observation that was not likely to be fraudulent with a 0. The Capstone team then linked this field to the file containing all the POS data.

The Capstone team reduced the sample size of the data used for regression proportionally to the number of claims that were reinvestigated by the SIU. The team assumed that the original data was representative of the following ratios:

- Observations without claims filed to observations with claims filed;
- Observations with claims filed to observations investigated by an SIU;
- Observations investigated by an SIU to fraudulent observations.

When the Capstone team reduced the sample size of the data used for regression, the team kept the above ratios the same for each model. Thus, the Capstone team felt that the data used for each model was representative of a random sample of the Auto-Insurance Company's customers.

Logistic Regression

Once the data ready was for regressions, it was entered it into a statistical analysis program called S-plus. The Capstone team used S-plus to run regressions and create predictive models. The fraud indicator from the reinvestigated claims was used as the response variable (y) and all the POS data as the predictor variables (x_i 's). Because the response variable was a 0 or 1, logistic regression was used to create the predictive models. The equation below represents a logistic regression model for 1 response variable (y) and k predictor variables (x_i 's).

$$E(y) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}$$

In the above equation, the β_i 's are coefficients for each predictor variable and $E(y) = \pi = P(y = 1)$ (Mendenhall & Sincich, 1996).

When the first logistic regression was run, there were over 100 predictor variables in the model. Predictor variables that did not significantly contribute to the model were removed. S-plus outputted statistical summary tables when each logistic regression was run. The Capstone team used the t-values in these tables to remove any predictor variables that did not statistically contribute to the model at an $\alpha = 0.05$ confidence level. After the statistically insignificant predictor variables were removed, the Capstone team looked at the signs of the coefficients (β_i 's) for each significant predictor variable. If the sign of a coefficient was opposite what the Capstone team expected it to be, the predictor variable was removed from the model.

After the removal of all the insignificant predictor variables, the Capstone team reran the regressions with only significant predictor variables to generate the final models. Once again, the team used the t-values and coefficients of each predictor variable to make sure that they were still significant. Then, the χ^2 values for the entire model were calculated to test the model's significance. If $(1 - \chi^2)$ equaled 1, it meant that the model was statistically significant.

MODEL DESIGN

The Capstone team wanted to create various models using different data sets and fields to predict fraud. Some of the data for the reinvestigated claims were incomplete. Some of the claims had complete data sets while other claims had incomplete data for fields that the team wanted to include in the model. The reinvestigated data also contained only claims from Florida. Although the majority of data came from Florida, the Capstone team wanted to make sure data from other states were included in some predictive models. Thus, the Capstone team created the following four models.

- **Model 1** contained complete data sets and a random sample of observations from all states.
- **Model 2** contained complete data sets and a random sample of observations from Florida only.
- **Model 3** contained incomplete data sets for the claims reinvestigated by the SIU and a random sample of observations from all states.

- **Model 4** contained incomplete data sets for the claims reinvestigated by the SIU and a random sample of observations from Florida only.

INDICES OF PERFORMANCE

When evaluating each model, the Capstone team wanted to minimize the statistical errors when the model was applied to a set of test data. For each model, the Capstone team removed a sample of data before running regressions. This data was used after the model was created to test the accuracy of the model. Specifically, the most accurate model was the one that minimized the Type I and Type II errors when applied to the test data.

In order to use the models the Capstone team created, the Auto-Insurance Company would choose a threshold probability above which, the potential customer would be charged a high premium or would be further investigated before being offered insurance. A Type I error is where the calculated probability is above the chosen threshold but the customer never committed fraud. A Type II error is where the customer committed fraud but the calculated probability is below the chosen threshold. The model with the lowest sum of Type I and Type II errors was chosen for proposal.

RESULTS

In order to measure the accuracy of each model, the Capstone team generated four graphs using each model against the test data. The Capstone team found Model 4 to be the most accurate.

Figure 1 below graphs the overall accuracy of Model 4. It shows the calculated probability that the observation would be fraudulent on the x-axis and actually whether or not the observation had a fraudulent claim on the y-axis. Ideally, all the fraudulent observations would have a high calculated probability of fraudulence while all the non-fraudulent observations would have a low calculated probability of fraudulence. This graph shows that there is a threshold probability where all observations with a higher calculated probability of fraudulence are actually fraudulent. This threshold probability is 0.70. Here, 28 of 51 fraudulent observations have a calculated probability of fraudulence higher than 0.70, while no non-fraudulent observations have a calculated probability of fraudulence higher than 0.70.

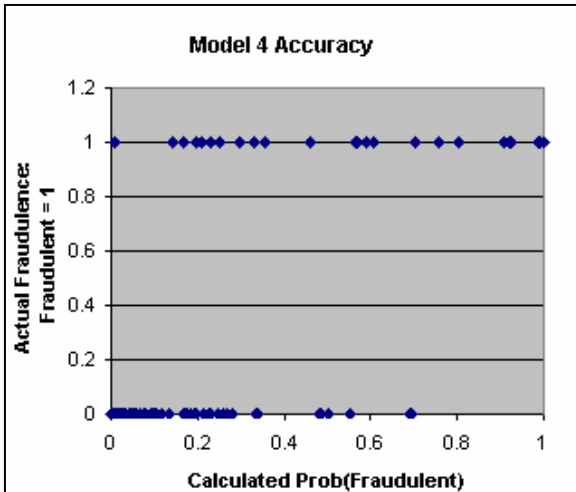


Figure 1 – Model 4 Accuracy: This chart shows that above a threshold probability of 0.70, 28 fraudulent observations have higher probabilities than the threshold while 0 non-fraudulent observations have higher probabilities than the threshold.

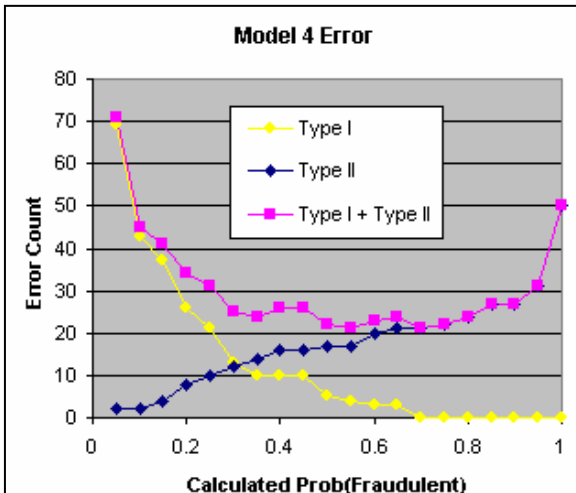


Figure 2 – Model 4 Error: This graph shows that the threshold probability that minimizes the error for Model 4 is 0.70.

Figure 2 above shows the calculated Type I and Type II errors that Model 4 generated when applied to the test data. It graphs the calculated probability that the observation would be fraudulent on the x-axis and the error count on the y-axis. Using this graph, the Capstone team determined that the threshold probability

where the total error is minimized is 0.70. Thus, this graph shows that Model 4 has the minimum error of all four models and that 0.70 is the threshold probability for Model 4 where the total error is minimized.

A Receiver Operating Characteristic (ROC) curve, shown in Figure 3 below, is a graph of the true positive rate versus the false positive rate based on various probability thresholds. It displays the tradeoff of losing true positives associated with minimizing the number of false positives (Tape 2002). The goal is to locate a point that both minimizes false positives and maximizes true positives. A curve that follows the left-hand border and then the upper border represents an accurate test. If the graph is located close to the 45° line, then the model should not be used for prediction. Lastly, the area under the curve describes the accuracy of the model. The accuracy of the model increases as the ROC curve bows further from the 45° line. Model 4's ROC curve is shown in blue in Figure 3. The black 45° line is a worthless model. This figure shows that Model 4 is accurate in predicting fraud.

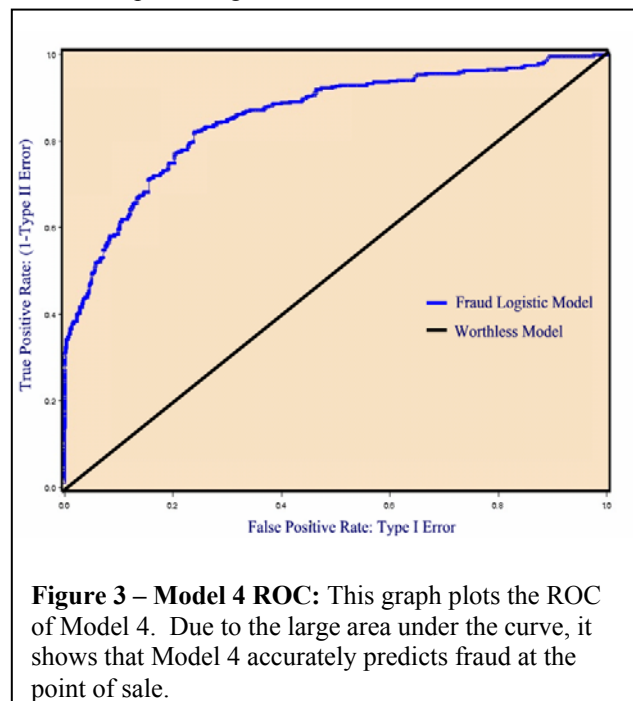


Figure 3 – Model 4 ROC: This graph plots the ROC of Model 4. Due to the large area under the curve, it shows that Model 4 accurately predicts fraud at the point of sale.

Figure 4 below illustrates the probability distributions of the fraudulent and non-fraudulent policies. It graphs the calculated probability of fraud on the x-axis and the observation count on the y-axis. The graph illustrates the mean calculated probabilities for fraudulent and non-fraudulent policies. It shows that the mean calculated probability for fraudulent

observations lies at 0.67 while the mean calculated probability for non-fraudulent observations lies at 0.03. This makes sense because the fraudulent policies have a high mean calculated probability of fraud while the non-fraudulent policies have a low mean calculated probability of fraud.

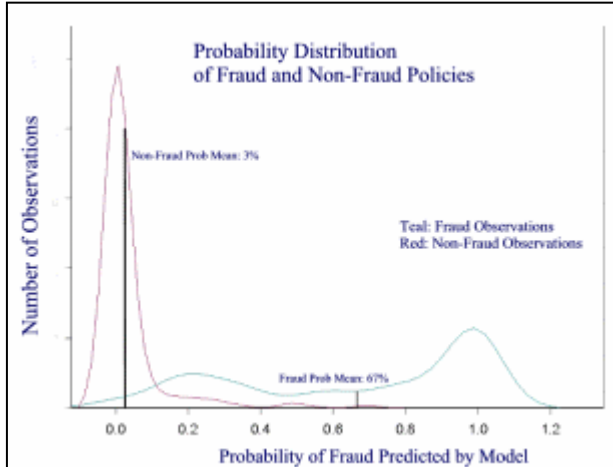


Figure 4 – Model 4 Probability Distributions: This graph shows the distribution of calculated probabilities of fraud for both fraudulent and non-fraudulent observations. The mean calculated probability for fraudulent observations lies at 0.67 while the mean calculated probability for non-fraudulent observations lies at 0.03.

The listed graphs from this section show the accuracy and usefulness of the proposed Model 4. Not only do these graphs show that Model 4 is statistically accurate, but they also show that the results generated from Model 4 make inherent sense. These graphs made the Capstone team confident that Model 4 will be accurate and have a strong impact on the Auto-Insurance Company.

IMPACT

The ability to recognize potentially fraudulent customers at the POS will be extremely beneficial to the Auto-Insurance Company. Most people who commit insurance fraud do so not only for personal benefits, but because it is easy to do. The proposed model has the potential to reduce the incentive to commit fraud by making it harder to do. When a potentially fraudulent customer tries to buy insurance, a very high premium will be charged reducing the fraudster's incentive to buy insurance. It will be harder

for fraudsters to get insurance, and thus, harder for them to commit fraud and collect on the insurance.

The elimination or suppression of fraud will minimize the Auto-Insurance Company's costs due to fraud throughout the year. The impact of minimized losses or maximized profits is two-fold. The financial impact is obvious, but the Auto-Insurance Company will also have the financial ability and motivation to lower premiums charged to customers. Customers will be trusted more because of the reduction of fraudulent claims and will expect lower rates. This decrease in rates will then increase customer clientele due to the more attractive policies, providing the Auto-Insurance Company with more business and the public with more affordable insurance.

Implementing the proposed fraud detection system at the Auto-Insurance Company could also be a risky investment. It is impossible to create a 100% accurate model, but an inaccurate model may discriminate against the innocent. A potential customer trying to buy insurance with no intention of committing fraud may be unfairly charged a high premium. Not only would the Auto-Insurance Company lose a perfectly good customer, but the customer could press charges against the Auto-Insurance Company, leading to a possible suit and court charges. Thus, it is paramount that the model be thoroughly tested before implementation.

The impact of this project is heavily dependent on the accuracy of the work done and of the proposed model. An accurate model will benefit the Auto-Insurance Company by increasing profits and reducing risk. An accurate system will also benefit customers by lowering premium rates. An inaccurate system could have a negative impact on the Auto-Insurance Company. After applying the proposed model to test data, the Capstone team feels confident that the project will have an overall positive impact on the Auto-Insurance Company.

The proposed system has the potential to change the strategy of all insurance companies. It may spread to competitors and the insurance industry as a whole. If this occurs, the impacts stated for the Auto-Insurance Company and its customers will multiply across the industry to increase the profits, will improve public perception of the insurance industry, and will allow consumers to achieve lower rates for an essential service.

FUTURE RECOMMENDATIONS

The proposed model is statistically significant and accurate according to the data used for modeling.

Before it is implemented, it is recommended that the Auto-Insurance Company test the model against completely new customers without applying the results. The results should be analyzed over the next year to determine the true accuracy of the model against actual POS customers.

The proposed model is also statistically significant and accurate for current fraud patterns. Fraud is a constantly changing crime. People who commit fraud change their schemes to perform each crime as society, business, and technology evolve. The proposed model could accurately predict fraudulent fire and theft claims now, but it could lose accuracy as people who commit fraud evolve their tactics. To ensure the accuracy of any model and the integrity of the company, the Auto-Insurance Company should constantly test any POS model for accuracy and adjust it to changes in fraud.

CONCLUSIONS

In completing this project, the Capstone team created a logistic model that is the first fraudulent predictive tool available to the Auto-Insurance Company. Using the proposed model, the Auto-Insurance Company can evaluate insurance applicants at the POS and determine the amount of financial risk they present to the company. The Auto-Insurance Company could create sales policies around an optimal probability generated by the proposed model, either elevating the premiums or completely denying coverage to extremely risky applicants. If the model proved successful in application, the Auto-Insurance Company could market predictive logistic tools to other insurance-based companies or credit-based industries.

REFERENCES

- Mendenhall, W. & Sincich, T. 1996. *A Second Course in Statistics: Regression Analysis*. Upper Saddle River, NJ: Prentice Hall.
- Tape, Thomas G. 2002. *Interpreting Diagnostic Tests*. Retrieved March 20, 2002 from the World Wide Web: <http://gim.unmc.edu/dxtests/Default.htm>
- Wright, J. 2001. Your Wheels; 'Victimless' Crimes That Hurt Us All. *Los Angeles Times*.

BIOGRAPHIES

Jeff Grossman is a fourth-year Systems Engineering student with a management concentration from Columbus, OH. Jeff will be working at Home Depot in the Business Leadership Program in Atlanta, GA next year.

Dawn Herndon is a fourth-year Systems Engineering student with a management concentration from Silver Spring, MD. Dawn will be working as a consultant at Booz, Allen, and Hamilton in McLean, VA next year.

Mark Michalski is a fourth-year Systems Engineering and Economics student from Brookfield, WI. Mark will be working at GE Medical Systems in the Information Management Leadership Program in Milwaukee, WI next year.

Carlton Pate is a fourth-year Systems Engineering student with a management concentration from Glastonbury, CT. Carlton will be working as a Systems Engineer at Applied Materials in Austin, TX next year.

Andrew Kaplan is a fourth-year Systems Engineering and Economics student from Sun Valley, ID. Andrew will be working as a Bond Risk Analyst at BlackRock in New York, NY next year.