Teaching Case

Alpha Insurance: A Predictive Analytics Case to Analyze Automobile Insurance Fraud using SAS Enterprise Miner™

Richard McCarthy
Richard.McCarthy@quinnipiac.edu

Wendy Ceccucci
Wendy.Ceccucci@quinnipiac.edu

Computer Information Systems
Quinnipiac University
Hamden, CT, 06518 USA

Mary McCarthy
Mary.McCarthy@ccsu.edu

Accounting Department
Central Connecticut State University
New Britain, CT, 06050, USA

Leila Halawi
halawil@erau.edu

Management Information Systems
Embry-Riddle Aeronautical University
Daytona Beach, FL, 32114, USA

Abstract

Automobile Insurance fraud costs the insurance industry billions of dollars annually. This case study addresses claim fraud based on data extracted from Alpha Insurance's automobile claim database. Students are provided the business problem and data sets. Initially, the students are required to develop their hypotheses and analyze the data. This includes identification of any missing or inaccurate data values and outliers as well as evaluation of the 22 variables. Next students will develop and optimize their predictive models using five techniques: regression, decision tree, neural network, gradient boosting, and ensemble. Then students will determine which model is the best fit providing consideration of the misclassification rate, average square error, or receiver operating characteristic (ROC). Lastly, students will generate predictive scores for the claims and evaluate the result using SAS Enterprise Miner™. Ultimately, the goal is to build an optimal predictive model to determine which of the automobile claims are potentially fraudulent.

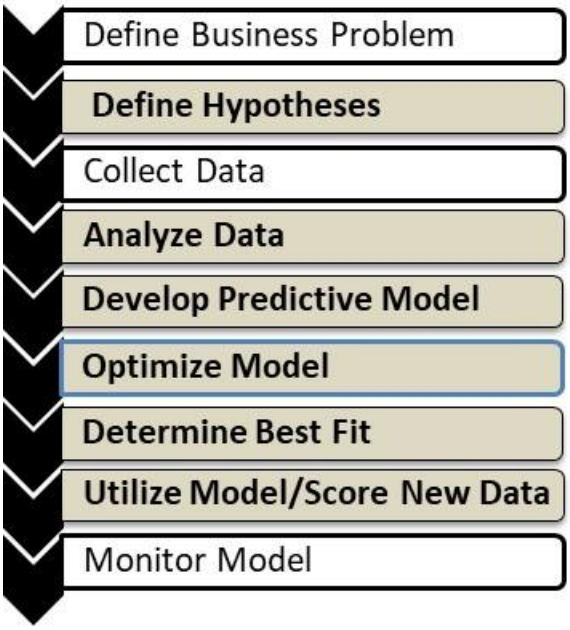
Keywords: predictive analytics, neural network, decision tree, regression, data mining, predictive scores, SAS Enterprise Miner

1. INTRODUCTION

This case is designed to be used in a predictive analytics course. The case provides an opportunity for extensive research and analysis of six of the nine steps in our Predictive Analytics Process Model (see Figure 1). Predictive techniques in the case include the *Big Three* - regression, neural networks, decision trees as well as Bayesian networks

Students are provided the business problem as well as the data. The business problem is to determine which new claims have the highest probability of fraud. However, based upon the data provided, the students must determine which hypotheses will be the focus of their analysis. They must then analyze the data and create their initial predictive model. Once the model is constructed, they can then optimize each node to determine the best fit. Finally, the new data can be scored from the best fit to determine the new claims that have the highest probability of fraud.

Figure 1 Predictive Analytics Process Model



Background

Despite recent developments in data analytics techniques and technology, the cost of fraud to the insurance industry continues to increase globally. According to the Coalition Against Insurance Fraud (2018), at least $80 billion is stolen each year as a result of insurance fraud.

Fraud is a common and recurrent problem in the property-casualty insurance industry. Insurers must be vigilant in identifying and dealing with fraudulent claims. Claim fraud analysis is a key analytic for many property-casualty insurers and most have a dedicated Special Investigative Unit (SIU) to investigate and resolve potentially fraudulent claims (Saporito, 2015). According to the Insurance Information Institute (2018), 42 states and the District of Columbia have set-up fraud bureaus for reporting potentially fraudulent claims. In some cases, they have multiple bureaus by line of business. Healthcare, workers compensation, and automobile insurance have been the three most prevalent lines of business to experience fraudulent claims. Insurance fraud continues to be a big challenge for the industry, regulatory authorities, and the public worldwide. Data driven fraud detection offers the possibility of utilizing a massive volume of prior claim history to determine patterns that uncover new potentially fraudulent claims which can then be investigated. This can provide both a cost and workload efficiency (Baesens, Van Vlasselaer, Verbeke, 2015).

Some activities that are fraudulent include vehicle dumping (i.e., the owner abandons or dumps the vehicle and reports it stolen), or exaggerated costs of repairs after an accident (Essurance, 2018).

Some of the techniques to predict insurance fraud include regression, neural networks, decision trees as well as Bayesian networks. Applications such as SAS Enterprise Miner™, IBM Watson Analytics, and Microsoft Power BI are used by insurance companies to help detect, analyze and ultimately reduce fraudulent activities. There are many tools and techniques in use to predict potentially fraudulent claims, therefore it is appropriate to use multiple techniques when analyzing specific claims.

The Case

The Alpha Insurance Company (this is a pseudonym and not intended to reference a specific organization) has contracted with you to develop an optimal predictive model to determine which of their automobile claims are potentially fraudulent. Historical data is a very good indicator of potential fraudulent claims, so it is appropriate to use it for analysis. They have provided two datasets for analysis. The first is a historical sample of automobile claim data containing 5,001 records. It contains attributes that are considered significant in the identification of

fraudulent claims, though it will be up to you, the analyst, to determine which of these attributes are the best determinants. The second file contains 4,008 current automobile claims that have not yet been analyzed. This file will be used to apply your best model to analyze which claims have the highest probability of being fraudulent (i.e., this is the dataset to be scored). This provides an opportunity to utilize the best predictive model to analyze current data.

At a minimum, Alpha Insurance would like you to utilize regression, decision trees and neural network models to determine the best model to predict which future claims are potentially fraudulent. These models are considered the *big three* in predictive analytics. In addition, you should consider gradient boosting and ensemble models.

The subsequent sections outline the requirements for each of the six required steps from the Predictive Analytics Process Model.

2. DETERMINE HYPOTHESIS

The data set used for this analysis contains 22 variables that each represent a single automobile claim.

Since the data source denotes the initial point for higher-level business analytics, data cleansing and data pre-processing efforts should be used.

Two of the variables are redundant and therefore may be rejected. These are the State (which is the expanded definition of the State_Code) and the Monthly_Premium (which is 1/12th of the Annual_Premium). Nine other variables are useful in understanding the cause and impact of the claim, however, they are not indicators of whether a claim is potentially fraudulent and thus should not be included in the analysis. They include: Vehicle_Model, Annual_Premium, Claim_Cause, Months_Since_Policy_Inception, Months_Since_Last_Claim, Claim_Report_Type, Location, Claim_Date, and Outstanding_Balance. The target variable is the Fraudulent_Claim indicator. This is a binary variable that documents whether the claim was fraudulent. It contains a value of Y (Yes) or N (No).

Appendix A provides a description of each of the attributes for both the sample historical data and as well as the current (score) dataset. From the remaining variables, you must then determine your hypotheses that is the subject of your analysis. Consideration must be given to whether all remaining variables will be subject to analysis or if additional variables will be rejected.

3. ANALYZE DATA

The sample claim data was extracted from Alpha Insurance's claim database.

Before you begin analysis, make sure your data source matches the roles and levels as described in Appendix A. The data needs to be processed to determine if there are any missing or inaccurate data values. In addition, outliers may have a significant impact on analysis and therefore they will also need to be considered. Alpha Insurance is interested in determining which factors (variables) are the most likely indicators that a claim is potentially fraudulent and what is the likelihood that the claim is fraudulent.

When preparing the data, you should test for outliers and missing values and handle them appropriately. You should also evaluate each of the independent variables to determine if any variables are skewed. If so, use appropriate transformations.

For your analysis, begin by partitioning the data using a 60/40/0 data set allocation for training, validation, and testing. Varying the partition sizes can impact the performance of a model. For a dataset of this size, it is possible to evaluate your models without creating a test dataset, later you may want to experiment with these settings.

4. DEVELOP AND OPTIMIZE PREDICTIVE ANALYTIC MODELS

Based upon the requirements set forth by Alpha Insurance, at least five techniques must be modeled to analyze this data (regression, decision tree, neural network, gradient boosting, and ensemble). For some of these techniques, it is appropriate to try several different approaches.

When performing a regression analysis, you should try several methods to determine which of these is the best fit model. These regression methods should include linear and/or logistic, multi-factor polynomials, and DMINE. When performing regression, consider the impact of utilizing stepwise, backward, or forward regression.

Decision trees are machine learning techniques that state independent variables and a dependent variable in a tree-shaped structure. Decision trees can vary in complexity, therefore when establishing your tree investigate the impact of

changing the depth and number of branches. Limit your depth to six and your branches to five to ensure that the tree does not have too many splits and therefore is no longer appropriate to explain the business problem.

Neural Networks vary greatly based upon the network type and number of hidden layers. Since we have a target variable to analyze, try both a generalized linear model and multi-layer perceptron model. Investigate the impact of varying the type of activation and combination functions as well as varying the number of hidden layers between two to six.

Each of the above techniques may result in the use of multiple nodes. Each node tested should be included in the final analysis. However, only the optimal node within each technique should be utilized within the ensemble node. If multiple partitions were tested, the results of the best performing partition should be considered in the final analysis.

5. DETERMINE THE BEST FIT AND SCORE NEW DATA

Alpha Insurance has not specified a specific selection statistic to be used as the basis for a recommendation on the model that is the best fit. Therefore, it is appropriate to consider whether the misclassification rate, average square error, or receiver operating characteristic (ROC) should be utilized, particularly if they yield a different result to determine which model is the best fit.

Once you have determined which model is the best fit, use that model to score the supplemental claim score data set to generate probabilities that these claims are fraudulent.

6. FINAL REPORT

The best fit model enables an insurance company to identify and detect potentially fraudulent activity more accurately and quickly, to ultimately reduce the payout on fraudulent claims.

In your final report, you must include the following sections:

1. Determine Hypotheses:
   What were the hypotheses that you tested? If any variables were excluded, discuss why they were removed from the subsequent analysis.

2. Analyze Data:
   Which variable(s) had missing values and how were they treated?
   Which variable(s) contained outliers and how did you address them?
   What variable(s) did you identify as being skewed and how did you handle them?
   What partition sizes were used and why?

3. Predictive Model:
   For each model type, document the properties that resulted in the best fit model?
   Which selection statistic was used and why? Show the results of all of the selection statistics.
   Which model type resulted in the best fit and why?

4. Scored Results:
   Which claimant number(s) had the highest probability of potential fraud and what were the probabilities?

7. REFERENCES

Baesens, B., Van Vlasselaer, V., Verbeke, W., (2015), Fraud Analytics: Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection, Wiley & Sons, Hoboken, NJ.

Coalition Against Insurance Fraud, (2018), Fraud: Why Worry? Retrieved on May 20, 2018 from www.insurancefraud.org/fraud-why-worry.htm

Essurance, (2018), retrieved on August 10, 2018 from https://www.esurance.com/info/car/5-examples-of-car-insurance-fraud

Insurance Information Institute, (2018), Background On: Insurance Fraud, Retrieved on Jan 1, 2018 from www.iii.org/article/background-on-insurance-fraud

Saporito, P., (2015), Applied Insurance Analytics, Pearson Education, Upper Saddle River, NJ.

8. NOTES

Two datasets accompany this case. They are: Claim Raw Data – containing 5,001 records that represent a historical analysis of fraudulent claims; and Claim Score Data – containing 4,008 records to be processed to determine which new claims have the highest potential for fraud.

Editor Note: Teaching Notes accompany this case, contact the authors

9. APPENDIX A – DATA DICTIONARY

Attribute	Role	Level	Definition
Claimant _Number	ID	Interval	Unique identifier assigned to each claim
State_Code	Input	Nominal	Two-letter state abbreviation where the claim occurred
State	Reject	Nominal	Name of the state where the claim occurred
Claim_Amount	Input	Interval	Total amount paid for the claim
Education	Input	Nominal	Level of education attained by claimant (High School or Below, College, Bachelor, Master, Doctorate)
Claim_Date	Reject	Nominal	Date when the claim occurred
Employment_Status	Input	Nominal	Employment status of the claimant (Employed, Unemployed, Medical Leave, Disability, Retired)
Gender	Input	Binary	Code indicating the claimant's gender (F, M)
Income	Input	Interval	Annual income of the claimant (in USD)
Location	Reject	Nominal	Categorical location where the claimant resides (Residential, Suburban, Urban)
Marital_Status	Input	Nominal	Marital status of the claimant (Divorced, Married, Single)
Monthly_Premium	Reject	Interval	Monthly premium amount for the policy
Annual_Premium	Reject	Interval	Annual premium amount for the policy
Months_Since_Last_Claim	Reject	Interval	Number of months since the last time the claimant had a claim prior to this claim
Months_Since_Policy_Inception	Reject	Interval	Number of months since the insured began policy coverage
Claim_Cause	Reject	Nominal	Cause of the claim (Collision, Fire, Hail, Other, Scratch/Dent)
Claim_Report_Type	Reject	Nominal	Code indicating how the claim was reported (Agent, Branch, Call Center, Web)
Vehicle_Class	Input	Nominal	Type of automobile damaged as a result of the claim incident (Two-Door Car, Four-Door Car, Luxury, SUV, Luxury SUV, Sports Car)
Vehicle_Size	Input	Nominal	Category indicating the size of the vehicle that was damaged (Compact, Midsize, Luxury)
Vehicle_Model	Reject	Nominal	Model of the vehicle that was damaged (Chevrolet, Ford, Honda or Toyota)
Outstanding_Balance	Reject	Interval	Remaining balance owed on the vehicle by the claimant at the time the claim occurred
Fraudulent_Claim	Target [Dependent]	Binary	Code indicating if the claim was fraudulent (Y/N)