

Large-Loss Predictive Model

Peter Fill
Zhongping Gao
Justin Warner

with the assistance of

Yiqun Hu

April 16, 2010

Abstract The goal of this report is to provide a model accurately predicting whether a worker's compensation policy will suffer a large loss. Several models were created using statistical analysis and the final logistic regression model predicting client policies at risk for large losses is provided. This model is a two-factor logistic regression model accurately predicted the large loss outputs of 87.5% of all policies and 70.1% of those policies suffering large losses. All data was supplied by the Actuarial Department of Accident Fund Insurance Company.

Table of Contents

Introduction.....	1
Methodology.....	1
Data Preparation	2
Models	3
Results.....	4
Discussion.....	6
Conclusions.....	8
Recommendations.....	8
Acknowledgements.....	9
References.....	10
Appendix I	11
Appendix II.....	13

Introduction

Accident Fund Insurance Company specializes in workers' compensation insurance. The features of a particular policy, such as the premium, are determined by the type of job performed by all employees and their predefined risk levels. Naturally, an employee with a high-risk level will require a more expensive policy, while an employee with a low-risk level will require a policy which is less expensive. The total premium for a policyholder is the sum of the premiums for all employees, and is determined in cost per one hundred dollars payroll. The average account sold by Accident Fund costs approximately \$1,000 per year.

Accident Fund Insurance Company would ideally pay for many small claims in the policies they sell. In addition to making a profit, this business model would generate data that could be used to refine the premium pricing models. Unfortunately, the unpredictability of the insurance industry ensures that the company will incur large losses, defined as payouts of more than \$100,000. Of course, the occurrence of a large loss can drastically reduce profits and hinder company operations.

The goal of this project is to create a predictive model to determine the probability that a certain policy will incur a large loss. This model will allow Accident Fund Insurance Company to revise the premiums on existing policies that are found to be at high risk for a large loss, and better determine the appropriate premium for new policies. As a result, large losses should be reduced and profitability should increase.

Methodology

Logistic regression was the primary method used to produce the predictive model. In statistics, logistic regression predicts the probability of an occurrence by fitting sample data to a logistic curve. Like many forms of regression analysis, it makes use of several predictor variables that may be either continuous or categorical (discrete).

An explanation of logistic regression begins with an explanation of the logistic function,

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}. \tag{1}$$

The domain of the logistic function is the entire real line and the range is confined between 0 and 1 [1]. Thus, the variable z represents the exposure to some set of independent variables, while $f(z)$ represents the probability of a particular outcome.

The variable z is a measure of the total contribution of all of the independent variables used in the model and is known as the logit. It is usually defined as $z = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n$ where the constant β_0 is called the intercept and the other constants β_k are called the regression coefficients of their respective parameter x_k . The intercept is

the value of z when the value of all independent variables is zero. In other words, the intercept is the value of z in a policy with no risk factors. Each of the regression coefficients describes the contribution of that risk factor. A positive regression coefficient means that that explanatory variable increases the probability of the outcome, while a negative regression coefficient means that variable decreases the probability of that outcome. Furthermore, the larger the magnitude of a regression coefficient is, the larger the influence of that risk factor on the probability of the outcome [2].

There are two types of logistic regression: binomial and multinomial. Binomial logistic regression is used to describe the relationship between one or more independent variables and a dichotomous variable that has only two possible values. The output is binary, rather than continuous, and is often used to predict death (the output is either “dead” or “not dead”). Multinomial logistic regression produces an output that is a continuous probability [1].

A critical component to the creation of a logistic regression model is the determination of the input variables. SPSS, the statistical software package used in data analysis, allows the user to choose between three different processes to determine input variables. The three methods are forward selection, backward elimination, and stepwise. While all techniques produce similar choices of input variables, it was determined that the stepwise method for variable selection would be used. The stepwise method is a modification of the forward-selection technique and differs in that variables already in the model do not necessarily stay there. As in the forward-selection method, variables are added one by one to the model. After this process, however, the stepwise method looks at all the variables already included in the model and deletes any variable that has the least significance, above the chosen significance level of 0.10, in determining $f(z)$. Only after this check is made and the necessary deletions accomplished can another variable be added to the model. The stepwise process ends when the variable to be added to the model is the one just removed [3].

Data Preparation

Detailed data on over 75,000 workers’ compensation insurance policies was provided from Accident Fund Insurance Company in Microsoft Excel spreadsheets.

Several procedures were taken to prepare the data for analysis. Some variables were immediately eliminated from the analysis because they contained duplicate information. All binary variables were coded numerically as 1 or 0 (Yes or No) to allow for complete analysis. The large loss indicator was additionally coded in this way. The government class code and policy hazard code were converted from their original alphanumeric values into numeric values. The codes A through F were converted to 0 through 6 respectively.

Finally, It was determined that state and industry were critical variables, however, the dimension of each was too large. Originally, there were 20 industries and 33 states. It was originally unclear that these inputs had significance and it was determined that they could be recoded to both reduce their dimension and increase their predictive power. Through a recoding process detailed in Appendix II, both state and industry were recoded and their dimension was reduced to two and six inputs respectively. The final conditioning of the data set was to remove all policies missing hazard codes, since they are a liability to the accuracy of the model.

For a complete list of variables, their definitions, and the process of their preparation for input please see Appendices I and II beginning on page eleven.

Models

In the process of determining a model to predict policies at risk for large loss claims several statistical methods were attempted. Many of these methods while informative, ultimately proved to inaccurate to be considered for the final model. Each model contributed to a new improved model or a complete departure from the method if the accuracy of the results were not satisfactory.

The first method attempted was Principle Component Analysis. This method converted the original twelve variables into a reformulated five variable scheme, each variable labeled as z_1, \dots, z_5 [4]. These new variables were linear combinations of the original twelve variables, created using SPSS software. Using the new five variables, a trend was extracted that adequately modeled approximately 60.2% of the data. In an attempt to improve the accuracy of the model generated using Principal Component Analysis the numerical values were standardized. For all continuous numerical variables the mean was subtracted and the result was divided by the standard deviation. The accuracy increased to 75.4% following standardization. This result seemed satisfactory, but it is unclear if another method may prove more effective.

The next model attempted was Logistic Regression (LR) model based on reconditioned variables. To redefine the variables, each variable was plotted against its propensity for a large loss and then rerated on a scaled determined subjectively. The first reconditioned variable was State. It seemed as though some states were more likely to suffer large losses than others, however after further analysis it was clear that these states suffered only from a relatively limited data set. It was determined that states with a total policy counter under a given threshold were grouped into an "Other" state. The LR model using a reconditioned state variable did predict, with 99.8% accuracy, the policies that did not incur a large loss. However, the model only predicted those policies that incurred a large loss with 10.0% accuracy. To improve this model the Policy Hazard Code was examined for reconditioning. It seemed as though a concentration of large losses occurred in policies near in the middle of the hazard code spectrum. Thus, Policy Hazard Code was subdivided into a low-risk, mid-risk, and high-risk level. The accuracy

of a new LR model formed with recondition Policy Hazard Code did not improve. Several threshold limits were tested for both the State and Policy Hazard Code variable was performed to find the optimal threshold for the new risk variable without a significant improvement in model accuracy.

The next refinement is to reduce the output data bias. Since 98.8% of policies in the original data set did not suffer a large loss. To correct this bias, five hundred policies that did not suffer a large loss were selected at random, and then modeled against five-hundred randomly selected policies that did suffer a large loss. This procedure was repeated several times with different random selections. The models generated from a balanced data set proved to be the most accurate. In early versions of this scheme, approximately 80.1% of the data is predicted accurately. Among policies that did not suffer a large loss, the accuracy was approximately 90.2%, and among those policies that did suffer a large loss, the accuracy was approximately 70.0%. Furthermore, when the probability of a policy suffering a large loss was calculated, a plot of probability percentile against the number of policies that suffered a large loss did increase as percentile increased. Due to the preliminary accuracy of this model, it was chosen to be further refined into a final product.

The final unbiased models were created using only two to four independent variables, and the selection of these variables was made for each random data set. In other words, for certain choices of data, certain parameters were more influential than others. Those variables that appeared often in these models were the model premium, the industry, the policy hazard code, the median days lost, and the state. Thus, it was determined that a new unbiased model should be created using these variables. In this particular case, only two parameters were recognized as significant: the model premium and the median days lost. SPSS was utilized to create the model, which was constructed as a function yielding the probability of the occurrence of a large loss. This model was formulated using a random balanced 2007 data set, and tested for accuracy over the entire 2008 data. This model accurately predicted 87.1% of the all 2008 policy large loss outputs and over the entire data set, 2007 and 2008, accurately predicts 87.5% of all policy large loss outputs. It was determined that the accuracy of this final two variable LR model was excellent and this model is chosen to be the final model.

Results

The final model chosen to predict policies at risk for large losses, those greater than \$100,000, is a two variable Logistic Regression model, (1). The variable was determined to be $z = 6.09(10^{-5})x_1 + 0.25x_2 - 1.702$, where x_1 is the model premium and x_2 is the median days lost. This model outputs a continuous set between 0 and 1 and it was determined that outputs below 0.5 are not at risk for large losses and outputs above 0.5 are at risk. Using this threshold and the above model, 87.5% of the large loss outputs for all policies are accurately predicted for the total data set, 2007 and 2008. However, since the model was built on the 2007 data set the model more accurately predicts the 2007

data set, 87.9%, but this improvement is minor in comparison to its accuracy of 87.1% over the 2008 data set. Also, since the data is heavily biased towards claims, which did not suffer large losses, the model was evaluated over only the policies that actually suffered large losses for both 2007 and 2008. This model accurately predicted 70.4% of these policies.

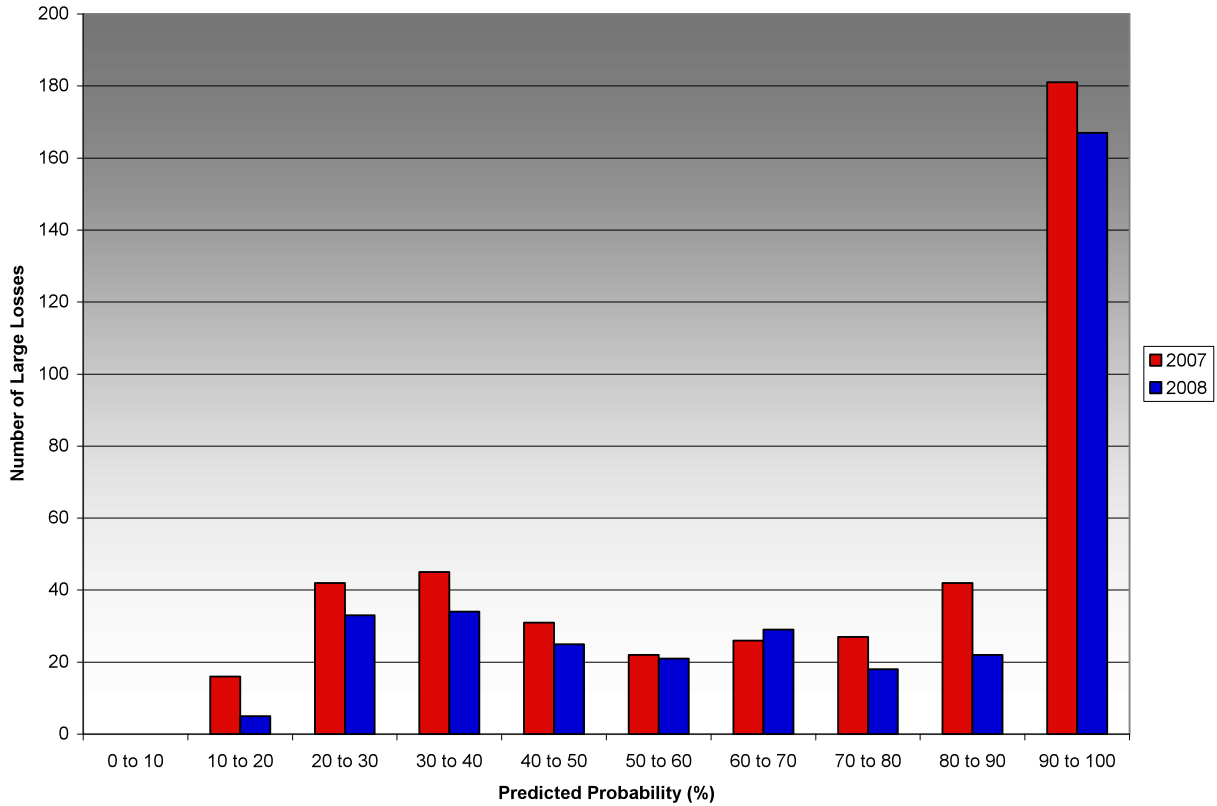


Figure 1. Histogram of predicted probability on policies that suffered large losses for the two variable LR model.

The above figure displays the probability predicted by the model only over policies, which suffered a large loss. This figure traditionally should appear as an exponential increase in probability, however it is most important that the bulk of the large loss policies appear in probability range from 50 to 100%.

The preceding LR model was created on a random set of balanced data from both 2007 and 2008 with 500 large losses and 500 not large losses. This model utilizes three independent variables and predicted the 2008 large loss outputs with an accuracy of 86.7%. While this accuracy is less than that of the two variable model it is still quite accurate. For this model, $z = 5.367(10^{-5})(\text{ModelPremium}) - 1.1(\text{IndustryThresholdA}) - 0.814(\text{IndustryThresholdB}) - 0.451(\text{IndustryThresholdC}) - 0.136(\text{IndustryThresholdD}) - 0.57(\text{IndustryThresholdE}) + 0.862(\text{StateThreshold}) - 0.875$, which is the input of (1). The accuracy of this model of the combined 2007 and 2008 data set is 86.7%.

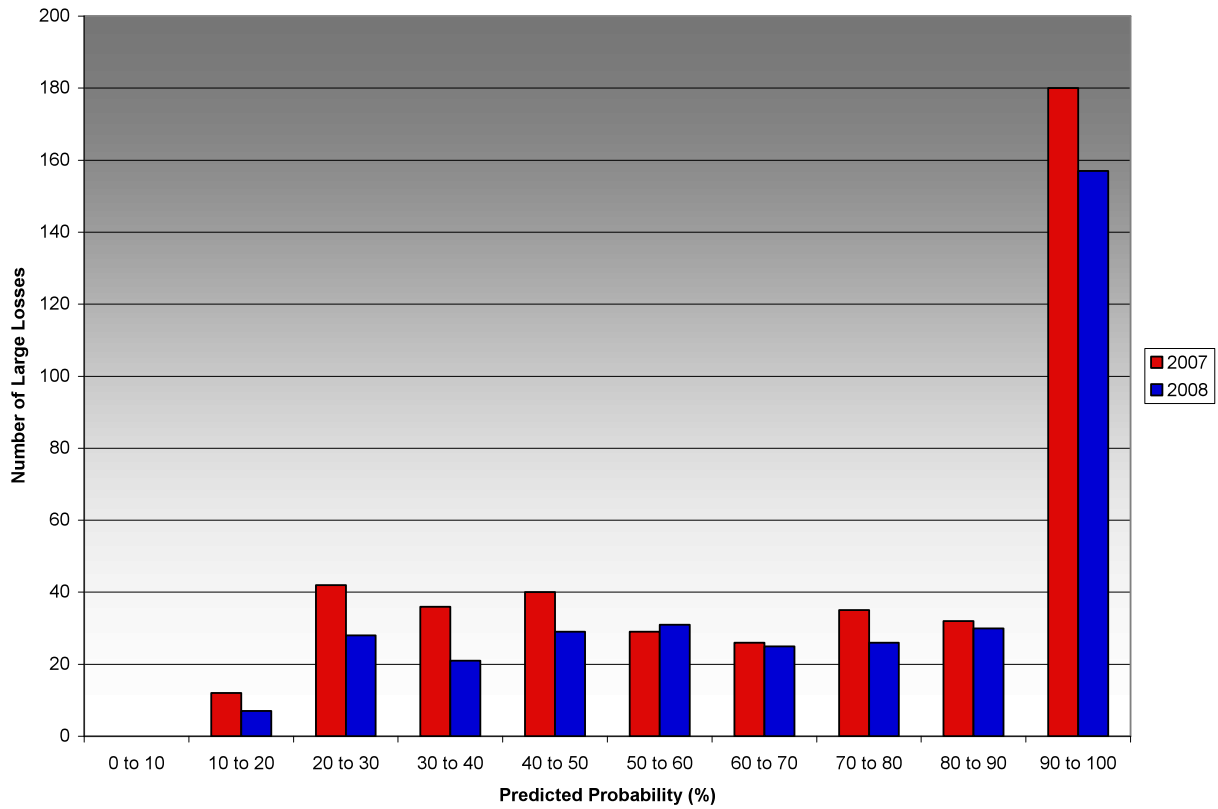


Figure 2. Histogram of predicted probability on policies that suffered large losses for the three variable LR model.

Notice that the three variable LR model behaves similarly to the two variable model in that both do not follow the traditional exponential curve but correctly predict a high level of the large loss policies over the 50 to 100% probability range.

Discussion

The first model was created using the technique of Principal Component Analysis, which transformed the original twelve variables into five new variables. This seemed reasonable as many of the variables were correlated. One specific example is the variable is the policy average rate, which was a function of total payroll and model premium (See Appendix I). The new five variable logistic regression model yielded an accuracy of 60.2% over the entire data set (2007 and 2008). Despite the 60.2% accuracy, the goal of the model is to accurately predict the large losses, in which the five variable LR model predicted only 10% accurately. While 10.0% accuracy does not make the model insufficient since the revenue saved by accurately predicting the 10.0% of those policies that did sustain a large loss is significant. The model would also inaccurately price other policies at higher levels causing other less risky policyholders unnecessary fiscal strain.

Depending on the pricing of these inaccurately predicted risky policies, the holders may consider finding an alternative insurance policy and the lost revenue to Accident Fund Insurance Company could be just as significant as that which was saved through accurate predictions. For this reason, new models must be considered.

Several Logistic Regression models were constructed, but major improvements in accuracy did not occur until a balanced data was used to construct the model. This conclusion was logical considering the original data set was heavily biased with policies that did not sustain large losses. The original data set contained 98.8% of policies not sustaining large losses. For this reason a 98.8% accurate model could have been constructed if the output was simply 0. This bias showed in the original Logistic Regression Models and was the main cause of the failure for each model to accurately predict large losses. These original models only accurately predicted around 10% of the large loss claims in the data set. However, once the set was divided into large loss claims and those not qualifying as large losses and a model was constructed from a balanced random sampling of these two sets did the accuracy improve considerably. The new models predicted 70.6% of the large loss claims, a 706% improvement over the unbalanced models. Overall, this model also accurately predicted 80% of the large loss outputs for the entire data set. The improved accuracy of large loss predictions and the relatively high accuracy over the complete data set led to a choosing a refined Balanced Logistic Regression model as a final result.

The final refined Balance Logistic Regression was created using a random balanced set of 600 data points from the 2007 data set. The significant factors were found to only be model premium and median days lost. This two factor LR model accurately predicted 70.6% of the large losses and 87.5% of all large loss outputs. It was initially surprising that just these two factors were capable of predicting 87.5% of the data. However, upon further review it appears that these two factors contain a significant amount of information about each policy. Firstly, model premium is defined by a complex proprietary formula, likely considering most if not all of the original variables given in the data set. Furthermore, the premium should indicate the riskiness of the given policy, since premium is priced based the expectation of the magnitude of the losses of the policy plus marginal profit for the insurance company. The second variable with predictive power is median days lost. This variable describes the median days lost due to a work injury by state and industry. There is clearly a correlation to both state and industry reflected in this variable. It is also important to note that the model created prior to the final model, which was generated from a balanced data set over both years, found the significant factors to be Industry, State, and Model Premium. Both models considered Model Premium significant, however the final model replaced industry and state with median days lost, which was previously described as an amalgamation of the two.

Despite that the final model uses just two variables to predict the probability a policy will suffer a large loss, it is clear that much more information is encoded into each variable and the accuracy of the model reflects this.

Conclusion

In closing, the probability of a worker's compensation policy to suffer a large loss, greater than \$100,000, is accurately modeled by a Logistic Regression model. The following conclusions are supported by the results of this study:

- Several models were created to model a policies large loss probability.
- The final two variable LR model accurately predicted 87.5% of the data.
- The model correctly predicted 70.6% of the actual large losses in 2007 and 2008.
- The two variables which demonstrated the most predictive power for a large loss are model premium and median days lost.

Recommendations

Logistic Regression and Principle Component Analysis were used to create several models attempting to predict the propensity of a large loss. However, these are not the only methods that can be used to construct such a model. Possible models to be used in future large loss analyses could include:

- Creating a Decision Tree to choose significant variables.
- Linear Discriminant Analysis, Multidimensional Scaling, or Perceptron to construct the probability model.
- Possibly combine these methods to create a model that has superior predictive accuracy.

These recommendations can be implemented in future work based on this report.

Acknowledgements

- Tower, Jack Mr. Jack Tower is a Predictive Modeling Analyst at Accident Fund Insurance Company. We would like to thank him for attending weekly progress report meetings and providing guidance throughout the formation of this report.
- Ferguson, Dana Mr. Dana Ferguson is an Actuarial Analyst at Accident Fund Insurance Company. We would like to thank him for attending weekly progress report meetings and providing guidance throughout the formation of this report.
- Ludden, Gerald Dr. Gerald Ludden is a professor of mathematics at Michigan State University and served as the Faculty Advisor to the report. We would like to thank him for attending weekly progress report meetings and keeping the report on track.
- Wu, Peiru Dr. Peiru Wu is a professor of mathematics at Michigan State University and the Director of the Master of Science in Industrial Mathematics Program. We would like to thank her for reviewing our reports and providing critical feedback.

References

- [1] Hosmer, David W., and Stanley Lemeshow. Applied Logistic Regression. Wiley series in probability and statistics. New York: Wiley, 2000.
- [2] Menard, Scott W. Applied Logistic Regression Analysis. Sage university papers series, no. 07-106. Thousand Oaks, Calif: Sage Publications, 1995.
- [3] SAS Institute Inc. 2008. SAS OnlineDoc® 9.1.3. Cary, NC: SAS Institute Inc.
- [4] Rijkevorsel, Jan L. A. van, and Jan de Leeuw. Component and Correspondence Analysis: Dimension Reduction by Functional Approximation. Wiley series in probability and mathematical statistics. Chichester [England]: Wiley, 1988.

Appendix I: Variables

Accident Fund supplied data on over 75,000 workers' compensation insurance policies in Microsoft Excel spreadsheets. Below are definitions and explanations of the supplied parameters.

- *Agency Tier*
 - This parameter indicates the level of Accident Fund's past loss experience for a given policy, valued as "Associate," "Select," or "Premier."
- *Deductible Flag*
 - This binary parameter indicates whether the policy is subject to a deductible.
- *Exponential Modification Factor*
 - This factor is multiplied to a policyholder's premium to adjust for past loss experience. The value is determined by Accident Fund.
- *Governing Class Code*
 - This categorical parameter is a numerical description of the occupational code corresponding to the job with the most annual payroll for a policyholder. Class Codes are a type of classification specific to workers compensation insurance.
- *Industry*
 - This parameter briefly described the industry sector of the policyholder.
- *Injury Illness Rate per 10K*
 - This continuous parameter produces the number of injury and illness cases reported for each ten thousand people for a given class code and state in the prior year. The rate is therefore not dependant on a specific policy, but rather the class code and state.
- *Large Loss Indicator*
 - This binary parameter states whether the policy suffered a large loss or not. A large loss is a loss of more than \$100,000.
- *Median Days Lost to Injury/Illness*
 - This continuous parameter gives the median number of days lost to injury and illness cases for a given class code and state in the prior year. The value is therefore not dependant on a specific policy, but rather the class code and state.
- *Model Average Rate*
 - This continuous parameter is constructed by the equation: $(100 * \text{Model Premium}) / \text{Total Payroll}$.
- *Model Premium*
 - This continuous parameter supplies the premium charged for a given policy.
- *New Business Flag*
 - This binary parameter indicates whether the policy has been written within the last calendar year.

- *Policy Hazard Code*
 - Originally a categorical parameter ranging from “A” to “G,” this code indicates the relative level of risk associated with the given policy, as determined by Accident Fund Insurance Company. The alphabetic values were converted to numerical values for analysis.
- *Renewal Counter*
 - This discrete parameter supplies the number of years the policy has been held with Accident Fund Insurance Company.
- *Short Term Flag*
 - This binary parameter indicates whether the policy duration is less than one year.
- *Total Payroll*
 - This continuous parameter gives the total payroll of the policyholder.
- *Unemployment Rate*
 - This continuous parameter yields the percent of unemployed individuals in the labor force in the county where the policy is issued.

Appendix II: Variable Conversions

Initially many of the variable inputs were required to be converted from categorical to discrete numerical outputs in order to be properly analyzed using the SPSS software suite. The following table details the treatment of each variable in the original data set.

Table 1. Variable Conversion Table.

Parameter	Type	Range	Conversion/Elimination
Agency Tier	Categorical: Text	NA	Eliminated
State	Categorical: Text	NA	Recoded
Deductable Flag	Binary	0 or 1	
Exponential Modification Factor	Continuous	-8.4 to 2.4	
Governing Class Code	Categorical: Numerical Code	NA	Eliminated
Industry	Categorical: Text	NA	Recoded
Injury Illness Rate per 10K	Continuous	10 to 629	
Large Loss Indicator	Binary	0 or 1	
Median Days Lost to Injury/Illness	Continuous	1 to 180	
Model Average Rate	Continuous	0 to 50,300	
Model Premium	Continuous	0 to 2,111,396	
New Business Flag	Binary	0 or 1	Eliminated
Policy Hazard Code	Categorical: Alphabetic	A to G	0 to 6
Renewal Counter	Discrete	0 to 26	
Short Term Flag	Binary	0 or 1	
Total Payroll	Continuous	0 to 86,630,480	
Unemployment Rate	Continuous	3.9 to 9.2	

For both state and industry, each variable option was measured as a sum of model premiums to determine whether there existed enough data within each subset. Variable options with low total model premiums were grouped into an “other” category to remove possible data imbalances. For the industry variable, the threshold for sum of model premiums was set to be 4.5 million, where all industry types containing less are regrouped into an “other” industry (See Figure 3).

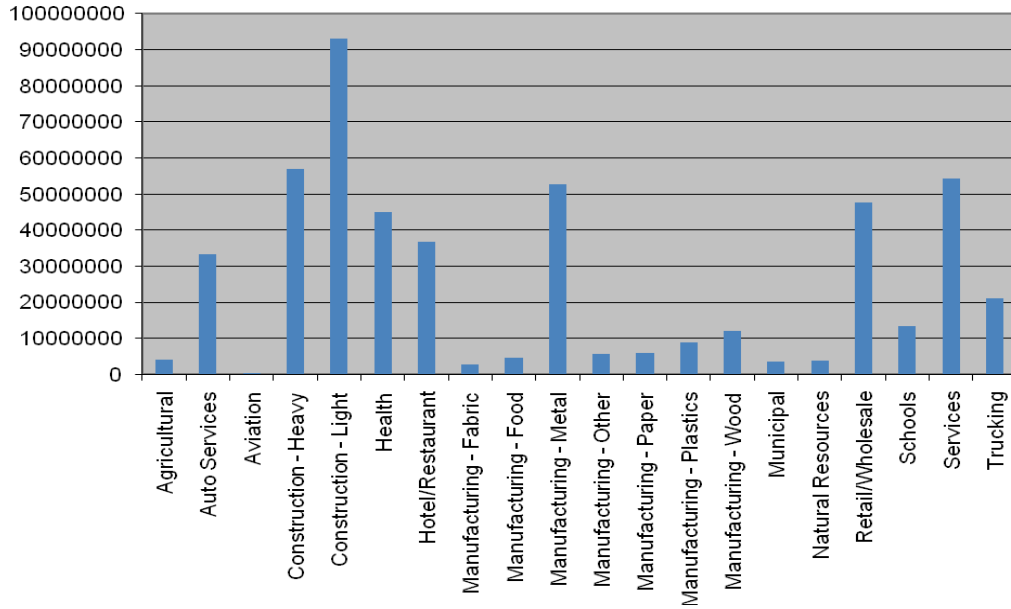


Figure 3. Sum of model premiums, in dollars, for each industry.

For the state variable, it was determined that all states with the sum of model premiums less than 1.7 million dollars are redefined as other (See Figure 4). For example, there is very limited data for the state of Louisiana and therefore all policies for the state of Louisiana are recoded to the state of “Other.”

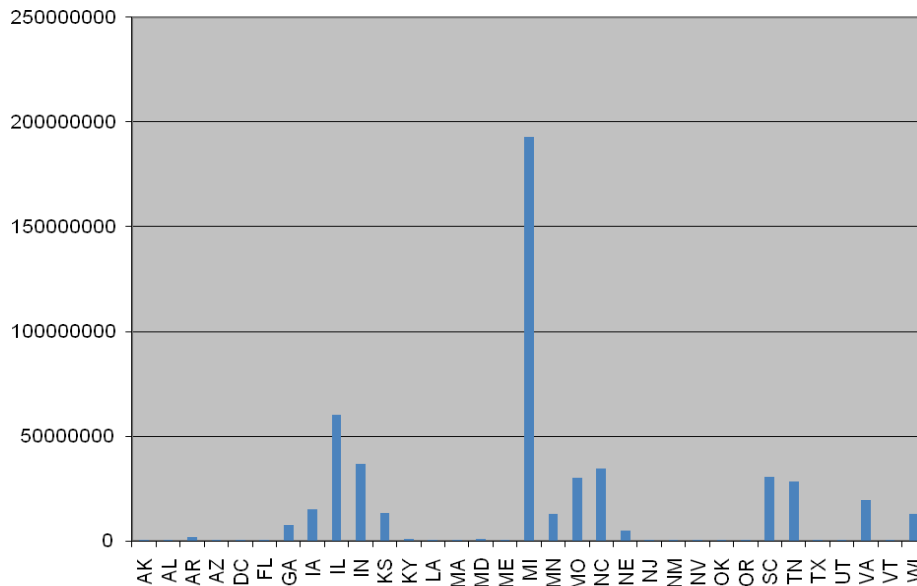


Figure 4. Sum of model premiums, in dollars, for each state.

Once both state and industry consisted of balanced inputs, they were then measured based on their propensity for a large loss. The probability of each industry to suffer a large loss was found in order to reclassify industries into a value based on

determined thresholds (See Figure 5). Based on Figure 5, six levels were found to classify the various industries into a industry risk variable. These levels were ranges of probabilities for large loss as seen in Table 2.

Table 2. Industry Risk Reclassification Table.

Range of Probability of Large Loss	Industry Threshold (Risk Level)
0-0.008	0
0.008-0.015	1
0.015-0.021	2
0.021-0.028	3
0.028-0.035	4
>0.035	5

For example, as seen in Figure 5, the manufacturing-paper industry had the highest probability of suffering a large loss and its industry threshold value was set to be 5.

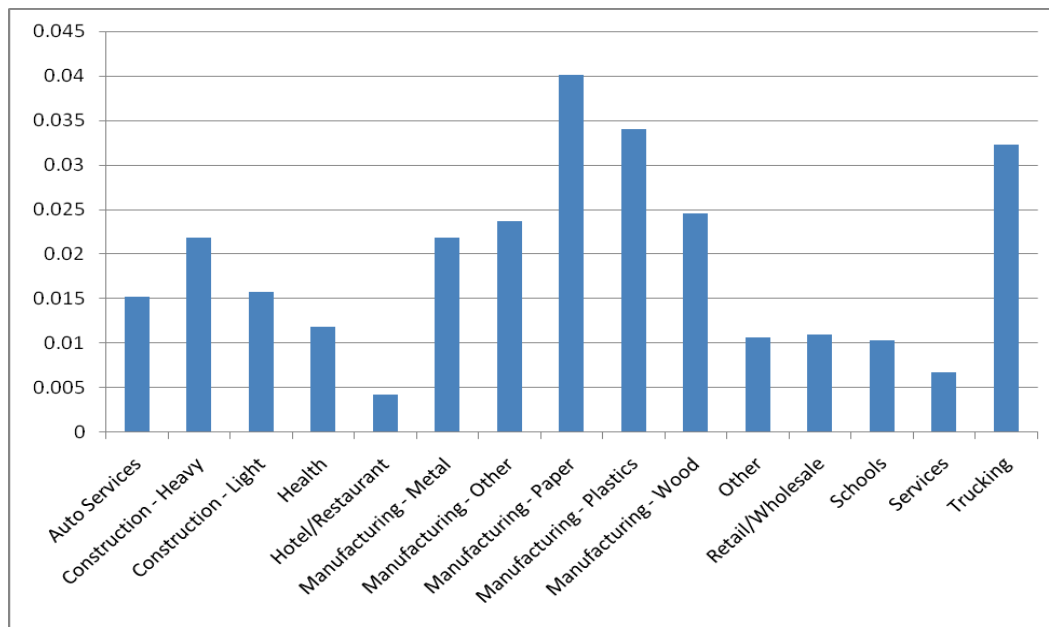


Figure 5. Percentage of policies suffering a large loss by industry.

A similar process was completed to reclassify the states following the creation of an “Other” state, however the final state threshold variable was binary, indicating a risky or non-risky state. The states determined to be risky were those with a probability for a large loss greater than 0.15, they were coded with a 1. States with a probability less than 0.15 were considered safer; these were coded with a 0 (See Figure 6).

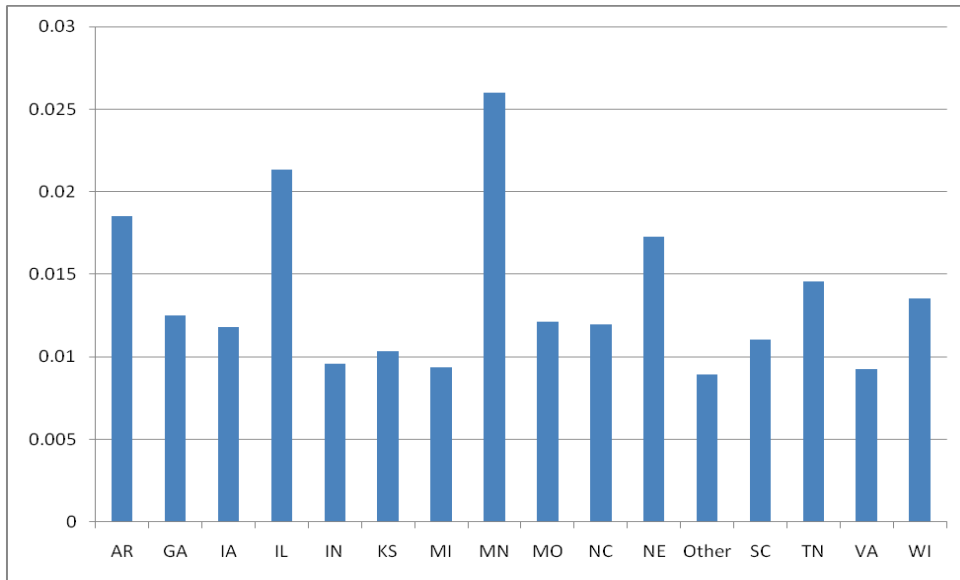


Figure 6. Percentage of policies suffering a large loss by state.