Claim Analytics Introduction: A Litigation Prediction Case Study

Mei Najim, Advanced Analytics Consulting Services, LLC. Chicago, IL

ABSTRACT

In the Property & Casualty Insurance industry, advanced analytics has increasingly penetrated into each of the core business operations – marketing, underwriting, risk management, actuarial pricing, and actuarial reserving, etc. Advanced analytics has been extensively used in claim areas to predict the claims' future in order to have early intervention to improve claim outcomes and claim processing efficiency. This is another important way to grow revenue and increase profit in the insurance industry.

This paper will first introduce several current common claim models in predictive analytics. Then, a claim analytics standard modeling process will be introduced. Finally, a Worker's Compensation litigation propensity model will be introduced as a case study. The steps in this litigation propensity modeling process will be discussed, which include business goal specification, model design, data acquisition, data preparation, variable creation/feature engineering, variable selection/feature selection, model building (a.k.a.: model fitting), model validation, model testing, model implementation, etc.

Base SAS®, SAS/STAT Enterprise Guide®, and SAS Enterprise Miner® are presented as the main tools for this standard process in this paper. Predictive modelers and data scientists have been using this process in R and Python as well. This process could also be tweaked or directly used in other industries, like healthcare, etc. This paper is intended not only for technical personnel but also for non-technical personnel who are interested in understanding a life cycle predictive model development general process.

INTRODUCTION

This paper has four parts. Part I provides an overview of the current common types of claim predictive analytics. Part II introduces a full life cycle of a predictive analytics standard modeling process from a business goal to model implementation and monitoring in claim analytics. Part III presents a Worker's Compensation litigation propensity model as a case study, in which each step on the modeling flow chart will be discussed in one separate sub-section. In Part IV concluding statements are made.

This paper provides readers some understanding about the overall claim predictive analytics modeling process and some general ideas on building models in Base SAS/STAT®, SAS Enterprise Guide®, and SAS Enterprise Miner®. Due to the proprietary nature of company data, some simplified examples with Census data have been utilized to demonstrate the methodologies and techniques which would serve well on large datasets in the real business world. This paper doesn't provide detailed modeling theory but instead focuses on application and this paper is good for all levels of audience, including technical or non-technical, especially in the P&C insurance industry.

AN OVERVIEW OF CLAIM ANALYTICS

In the claim industry, advanced analytics has been increasingly used and will be more extensively used in the near future (Please see the predictive modeling benchmark survey from Towers Watson).

Triaged claims can be assigned to the appropriate personnel and resources, claim settlement specialists can be optimized (even nurse assignments if necessary), and claim overall severity can be predicted in order to have early intervention to improve claim outcomes and claim processing efficiency. Claim analytics has been utilized not only for individual insurance companies but also for the large third party administrators, such as Sedgwick Claim Management Services and Gallagher Bassett Services.

Operation	2015 (Personal + Commercial)	2016		Two Years	
		Personal	Commerical	Personal	Commerical
Fraud Potential	28%	28%	66%	70%	55%
Claim Triage	17%	18%	15%	59%	66%
Litigation Potential	10%	23%	10%	54%	50%
Case Reserving	N/A	9%	8%	41%	48%

Source: Towers Watson – claim predictive modeling benchmark survey

Some Common Claim Predictive Models

Claim Assignment Model triages claims based on the nature of complexity to assign the claims to the appropriate personnel and optimize resources for claim settlement specialists

Clinical Guidance Model predicts if the claims need a nurse care management referral to guide a proper treatment

Complex or Large Loss Model predicts the claim overall severity to identify complex or large loss claims above certain thresholds to reduce the claim costs

Fraud Detection Model predicts and detects the possibility of presence of fraud in order to prevent fraudulent claims

Litigation Propensity Model predicts potential litigation to have early intervention and a settlement with claimants to avoid litigation and reduce litigation expenses.

Loss Reserve Model predicts the loss reserve amounts for claims when they are closed so that adequate reserve amounts could be booked properly and timely

Medical Escalation Model predicts claims that have a small medical incurred loss in the beginning but get escalated into large medical incurred loss

Recovery Model identifies claims with potential for successful subrogation, salvage and refund recoveries exceeding a threshold to improve recovery yield

Return to Work Model identifies claimants with a long return to work period to assist them to return to work earlier

Supervisor Focus Model predicts and identifies the claims that look small in the beginning but potentially develop into unexpected high severity claims

Near Real Time Claim Models

Open claims could have different characteristics at different ages as claims could keep developing until being closed. Claim adjusters (examiners) collect and enter required claim related information into the claim data system to help to book adequate reserve amounts and pay for the medical bills, etc. The goal of the claim handling process is to help injured workers to get appropriate and timely treatments to recover from injuries and return to work as soon as possible. An efficient claim handling process also helps insurance companies to reduce the expense to close the claims as early as possible.

To ensure the claim handling process is consistent and efficient across different adjusters (examiners) and different branch offices, there are usually a number of best practices with milestone time lines (as of claim day 30, day 45, etc.) for claim adjusters to follow to collect and enter required claim information. Given this current claim handling process standard practice and associated data availability, it makes sense to build near real time models corresponding with the associated milestone time lines, instead of real time models, in order to optimize cost and benefit for the claim handling unit.

A flowchart (Figure 1) for a near real time modeling process is shown below:

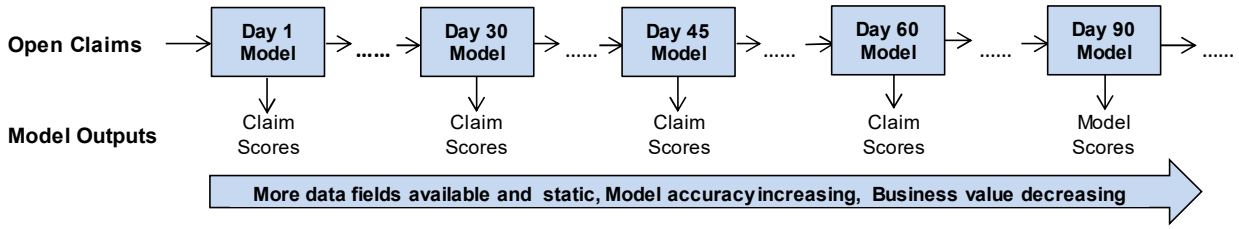


Figure 1. An Example of Near Real Time Modeling Process Flow Chart

A STANDARD PREDICTIVE MODELING PROCESS IN SAS

Any predictive modeling project process starts from a business goal. To attain that goal, data is acquired and prepared, variables are created and selected, and the model is built, validated, and tested. The model finally is evaluated to see if it addresses the business goal and should be implemented. If there is an existing model, we would also like to conduct a model champion challenge to understand the benefit of implementing the new model over the old one. In the flow chart on the next page (Figure 2), there are nine stages in the life cycle of the modeling process. The bold arrows in the chart describe the direction of the process. The light arrows show that at any stage, steps may need to be re-performed based on the result of each stage, resulting in an iterative process.

- ❑ Business Goals (and Model Design)
- ❑ Data Scope and Acquisition
- ❑ Data Preparation
- ❑ Variable Creation (a.k.a.: Feature Engineering)
- ❑ Variable Selection (a.k.a.: Feature Selection)
- ❑ Model Building (a.k.a.: Modeling Fitting)
- ❑ Model Validation
- ❑ Model Testing
- ❑ Model Implementation and Monitoring

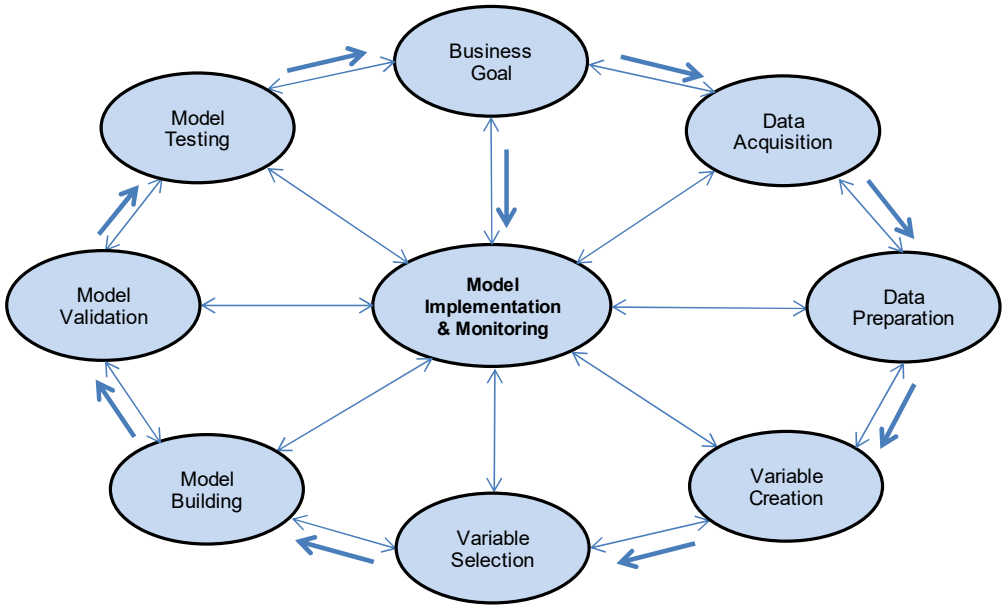


Figure 2. A Standard Property & Casualty Insurance Predictive Modeling Process Flow Chart

CASE STUDY – LITIGTION PROPENSITY PREDICTIVE MODEL

Over the past couple of decades, in the property & casualty insurance industry, around 20% of closed indemnity claims have been settled with litigation propensity representing 70-80% of total dollars paid. Indemnity claim severity with litigation is on average 4 to 5 times higher than the one without litigation. Apparently, litigation has been one of the main claim severity drivers. This model can predict which claims are likely results in litigation, and mitigate those claims to more senior adjusters who can settle the claims faster with lower costs.

In this case study, we are going to introduce a Litigation Propensity Predictive Model which is designed to predict the open claim litigation propensity in the future. The data includes WC Book of Business data with over a few thousands of clients' data. Multiple cutting-edge statistical and machine learning techniques (GLM Logistic Regression, Decision Tree, Gradient Boosting, Random Forest, Neural Network, etc.) along with WC business knowledge are utilized to discover and derive complex trends, patterns, and relationships across the WC book of business data to build the model.

1. BUSINESS GOALS AND MODEL DESIGN

The business goal is to predict and identify WC open indemnity claims with a high litigation propensity in order to prevent or settle the litigation to reduce claim severity and close the claims earlier. Based on the business goal, the model design is to build a model to predict and score the litigation propensity of open indemnity claims.

Therefore, the litigation flag is the dependent variable/target variable for the predictive model. Or, if a direct litigation flag is not available (which could happen to many companies with data systems which weren't designed for predictive models yet), a litigation flag proxy could be defined based on data knowledge and the business goal. In general, an ideal optimal litigation flag proxy should have achieved a targeted balance between the percentage of claims flagged as litigated and the percentage of associated total incurred loss

represented.

Usually, an ideal modeler/data scientist would have not only statistical and machine learning knowledge but also insurance claim data and business knowledge. Insurance claim data and business knowledge could take years to gain decent experience. To ensure the model is designed and supported by data and fulfill the business goal(s), each step of predictive modeling process and its findings should be presented and discussed with business and operation personnel on a regular basis until the model implementation.

2. DATA SCOPE AND ACQUISITION

Based on the specific business goals and the designed model, the data scope is defined and the specific data including internal and external data is acquired. Most middle to large size insurance organizations have sophisticated internal data systems to capture their exposures, premiums, and/or claims data. Also, there are some variables based on some external data sources that have been proved to be very predictive. Some external data sources are readily available such as insurance industry data from statistical agencies (ISO, AAIS, and NCCI), open data sources (demographics data from the Census Bureau), and other data vendors.

For this litigation propensity model, claim data, payment data, litigation data, managed care data, and other external vendors' data sources have been explored and used.

An example of data scope is: Coverage code = "WC"; Closed year between 1/1/2010 and 12/31/2016; Claim status="Closed", etc.

The rule of thumb: Try to keep consistency between the data used to build the model and the data that is going to be scored by the model. Even though there are some rare claims that are possible to randomly happen again, don't simply exclude them as we want to build a robust model.

3. DATA PREPARATION

3.1. Data Review (Profiling, Cleansing, Imputation, Transformation, and Binning, etc.)

Understanding every data field and its definition correctly is the foundation to make the best use of the data towards building good models. Data review is to ensure data integrity including data accuracy, consistency, and that basic data requirements are satisfied and common data quality issues are identified and addressed properly. If the part of the data isn't reflected the trend into future, it should be excluded.

For example, initial data review is to see if each field has decent data volume to be credible to use. Obvious data issues, such as blanks and duplicates, are identified, and either removed or imputed based on reasonable assumptions and appropriate methodologies. Missing value imputation is a big topic with multiple methods so we are not going to go into detail in this paper.

When we use 5 to 10 years of historical data, trend studies need to be performed and financial data fields need to be trended to adjust the year-to-year inflation and economic changes, etc.

Here are some common SAS® procedures in both Base SAS® and SAS Enterprise Guide®:

CONTENTS Procedure, FREQ procedure, UNIVARIATE Procedure, and SUMMARY Procedure.

The Data Explore Node in SAS Enterprise Miner® (Figure 3) is also a very powerful way to quickly get a feel for how data is distributed across multiple variables.

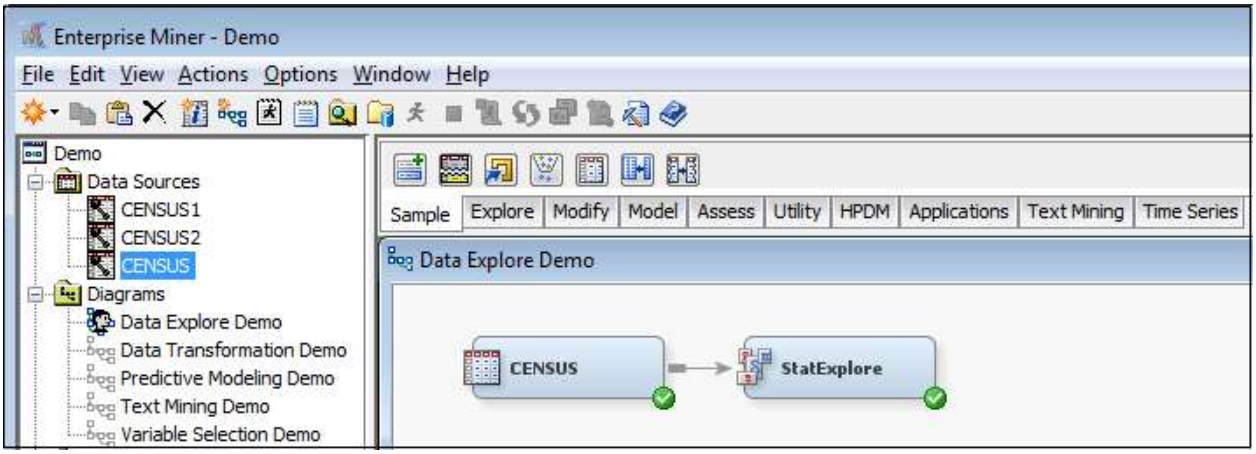


Figure 3. An Example of Data Review Using *CENSUS* and *StatExplore* Node in SAS ENTERPRISE MINER®

The diagram (Figure 4) below on the next page is the Data Explore feature from ***CENSUS*** node in the diagram (Figure 3) above. It shows how a distribution across one variable can be drilled into to examine other variables. In this example, the shaded area of the bottom graph represents records with median household income between $60,000 and $80,000. The top two graphs show how these records are distributed across levels of two other variables.

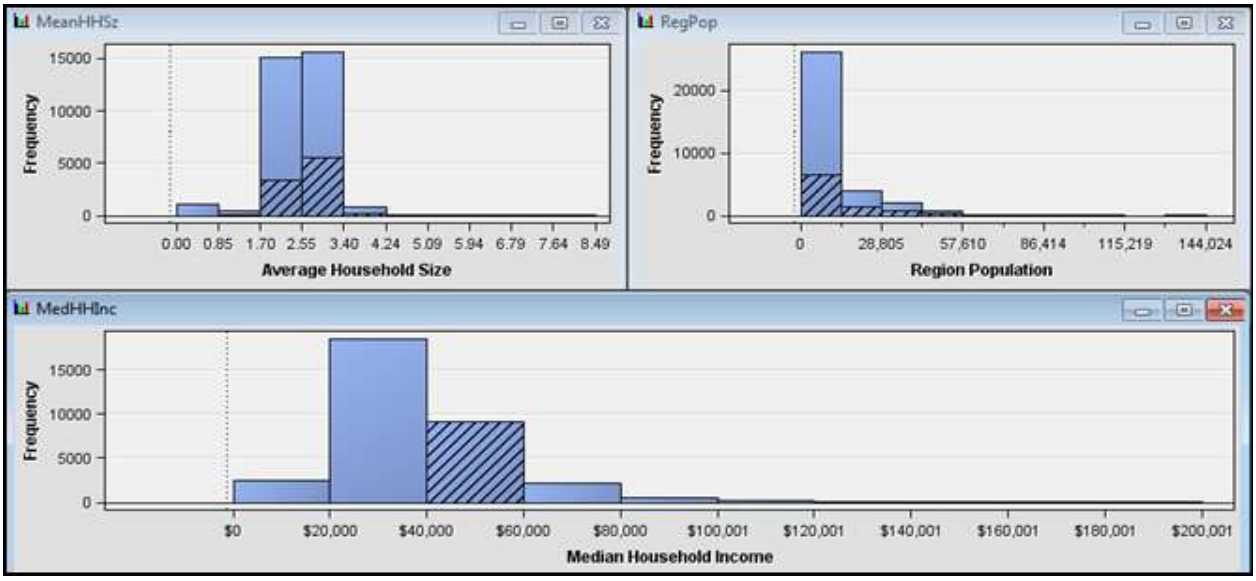
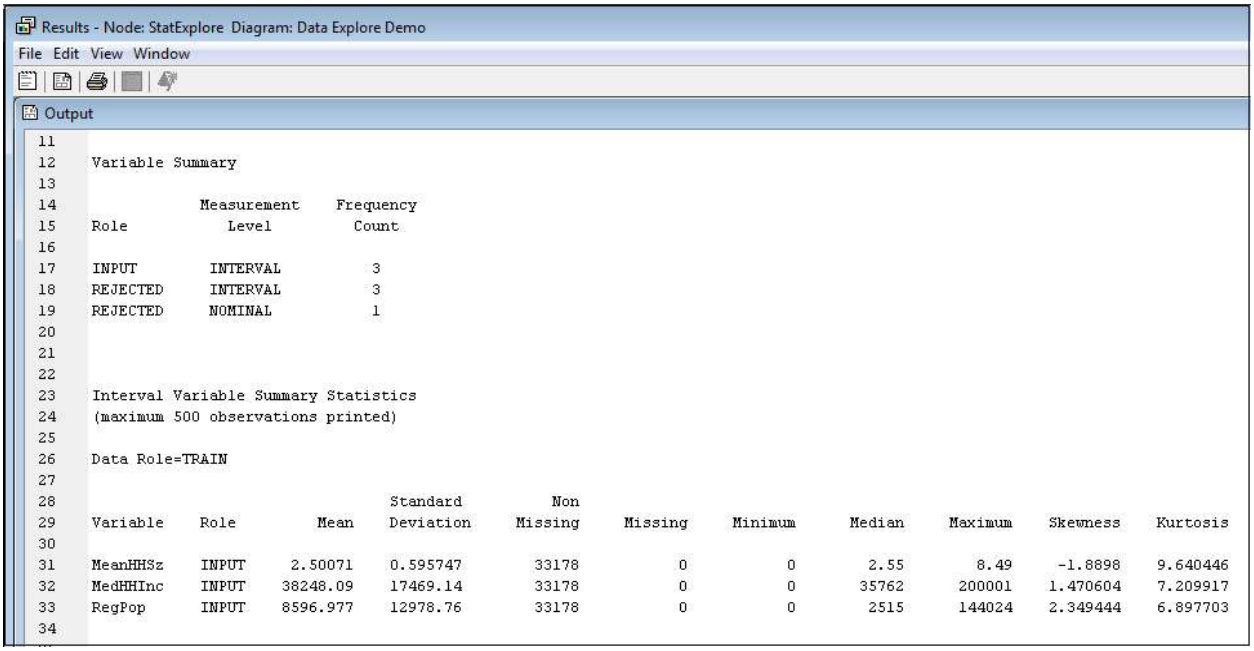


Figure 4. An Example of Data Explore Graphs from *CENSUS* Node in SAS ENTERPRISE MINER®

The result report below (Figure 5) with thorough statistics for each variable is provided after running ***StatExplore*** node in the diagram (Figure 3). The diagram (Figure 5) of a result report with three data fields from CENSUS data. When the data source contains hundreds or thousands fields, this node is a very efficient way to conduct a quick data explore.




```
Variable Summary

                Measurement    Frequency
Role               Level         Count

INPUT           INTERVAL           3
REJECTED        INTERVAL           3
REJECTED        NOMINAL            1



Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                  Standard          Non
Variable    Role          Mean    Deviation      Missing    Missing    Minimum     Median    Maximum    Skewness    Kurtosis

MeanHHSz    INPUT         2.50071   0.595747       33178          0          0       2.55       8.49     -1.8898    9.640446
MedHHInc    INPUT       38248.09   17469.14       33178          0          0      35762     200001    1.470604    7.209917
RegPop      INPUT       8596.977   12978.76       33178          0          0       2515     144024    2.349444    6.897703
```

Figure 5. An Example of Results Report Using *StatExplore* Node in SAS ENTERPRISE MINER®

When the variables exhibit asymmetry and non-linearity, data transformation is necessary. In **SAS ENTERPRISE MINER®**, ***Transform Variables*** node is a great tool to handle data transformation. The diagram below (Figure 6) is an example of data transformation procedure in **SAS ENTERPRISE MINER®**.

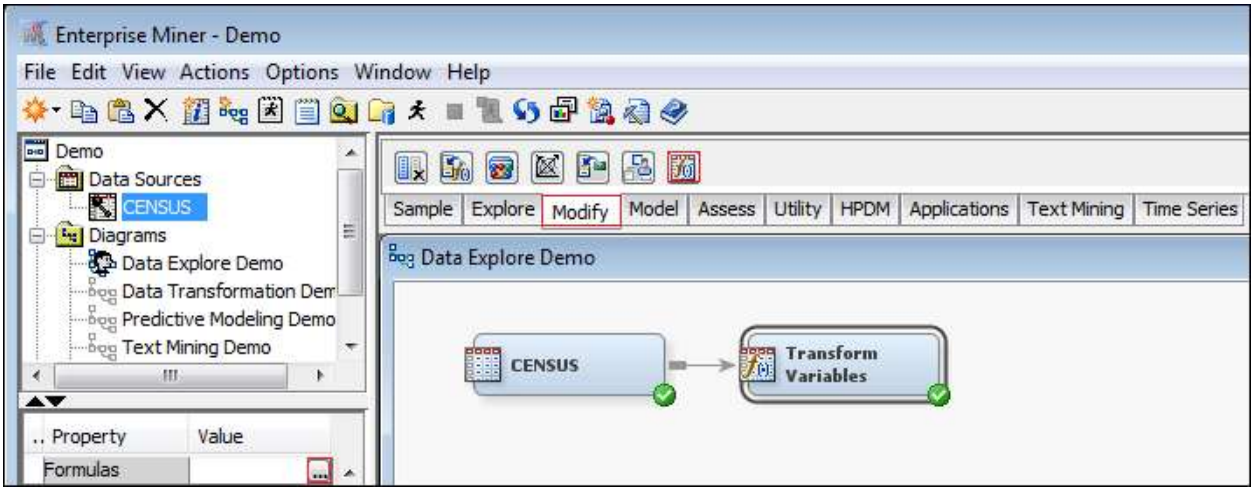


Figure 6. A Data Transformation Process Using *Transform Variables* Node in SAS ENTERPRISE MINER®

3.2 Data Partition for Training, Validation, and Testing

If data volume allows, data could be partitioned into training, validation, and holdout testing data sets.

The training data set is used for preliminary model fitting.

The validation data set is used to monitor and tune the model during estimation and is also used for

model assessment. The tuning process usually involves selecting among models of different types and complexities with the goal of selecting the best model balancing between model accuracy and stability. ***The holdout testing data set*** is used to give a final honest model assessment.

In reality, different break-down percentages across training, validation, and holdout testing data could be used depending on the data volume and the type of model to build, *etc.* It is not rare to only partition data into training and testing data sets, especially when data volume is concerned.

The diagram (Figure 7) below shows a data partition example. In this example, 80% of the data is for training, 10% for validation, and 10% for holdout testing. Other breakdown percentages could work too.

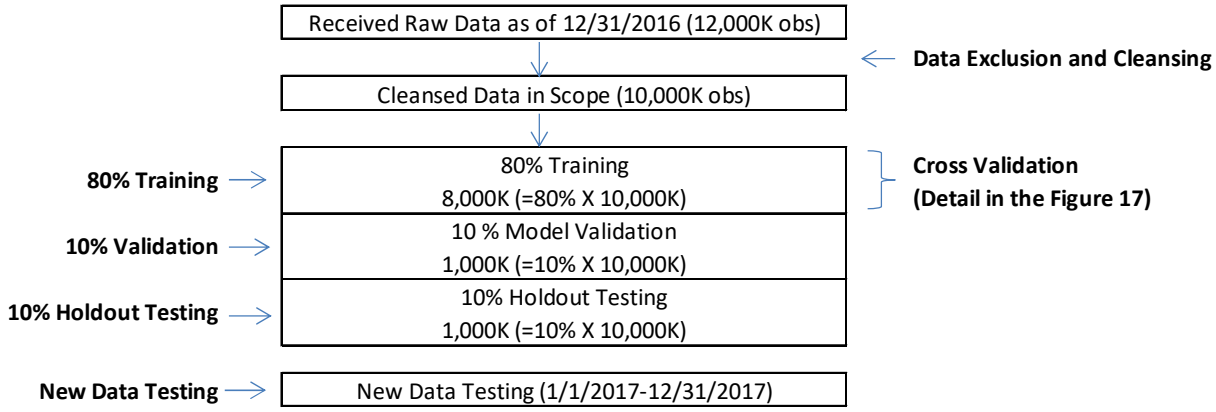


Figure 7. A Data Partition Flow Chart

The diagram below (Figure 8) is the data partition example in SAS ENTERPRISE MINER®. This node uses simple random sampling, stratified random sampling, or cluster sampling to create partitioned data sets.

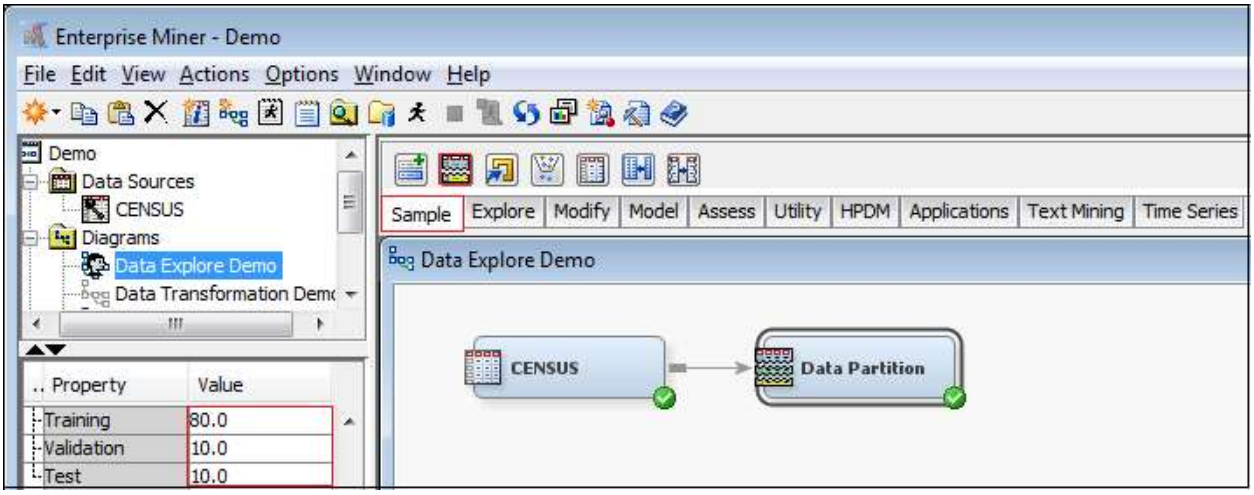


Figure 8. A Data Partition Procedure Example Using *Data Partition* Node in SAS ENTERPRISE MINER®

4. VARIABLE CREATION (a.k.a.: FEATURE ENGINEERING)

4.1 Target Variable Creation (a.k.a.: Dependent Variable or Responsible Variable)

Every data mining project begins with a business goal which defines the target variable from a modeling perspective. The target variable summarizes the outcome we would like to predict from the perspective of the algorithms we use to build the predictive models. The target variable could be created based on either a single variable or a combination of multiple variables.

For example, we can create the ratio of total incurred loss to premium as the target variable for a loss ratio model.

Another example involves a large loss model. If the business problem is to identify the claims with total incurred loss >= $250,000 and claim duration >= 2 years, then we can create a target variable = “1” when both total incurred loss >= $250,000 and claim duration >= 2 years, else “0”.

4.2 Predictive Variables Creation

Many variables can be created directly from the raw data fields they represent. For example: Number of dependents, Gender, Marital, Age, etc.

Other additional variables can be created based on the raw data fields and the business meanings, such as Prior claim indicator, Doctor payment fee binning, and Median household income ranking, etc. Example: Loss month can be a variable created based on the loss date field to capture potential loss seasonality. It could be a potentially predictive variable to an automobile collision model since automobile collision losses are highly dependent on what season it is. When the claim has a prior claim, we can create a prior claim indicator which potentially could be predictive variable to a large loss model.

4.3 Text Mining (a.k.a.: Text Analytics) to Create Variables Based on Unstructured Data

Text Analytics uses algorithms to derive patterns and trends from unstructured (free-form text) data through statistical and machine learning methods (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), as well as natural language processing techniques. The diagram (Figure 9) below shows a text mining process example in SAS ENTERPRISE MINER®.

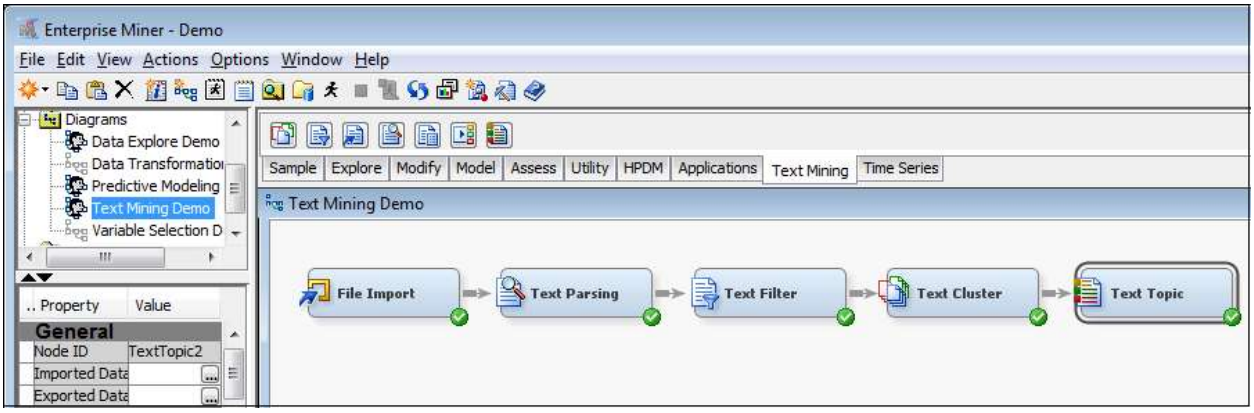


Figure 9. A Text Mining Procedure Example in SAS ENTERPRISE MINER®

4.4 Univariate Analysis

After creating target variable and other variables, univariate analysis usually has been performed. In the univariate analysis, one-way relationships of each potential predictive variable with the target variable are examined. Data volume and distribution are further reviewed to decide if the variable is credible and meaningful in both a business and a statistical sense. A high-level reasonability check is conducted. In this univariate analysis, our goal is to identify and select the most significant variables based on statistical and business reasons and determine the appropriate methods to group (bin), cap, or transform variables.

The UNIVARIATE procedure could be utilized in Base SAS® and SAS Enterprise Guide®. Some of the data review methods and techniques in data preparation could be utilized as well.

In addition to the previously introduced procedures, the diagram below (Figure 10) shows how Graph Explore procedure can also be used to conduct univariate analyses in SAS ENTERPRISE MINER®.

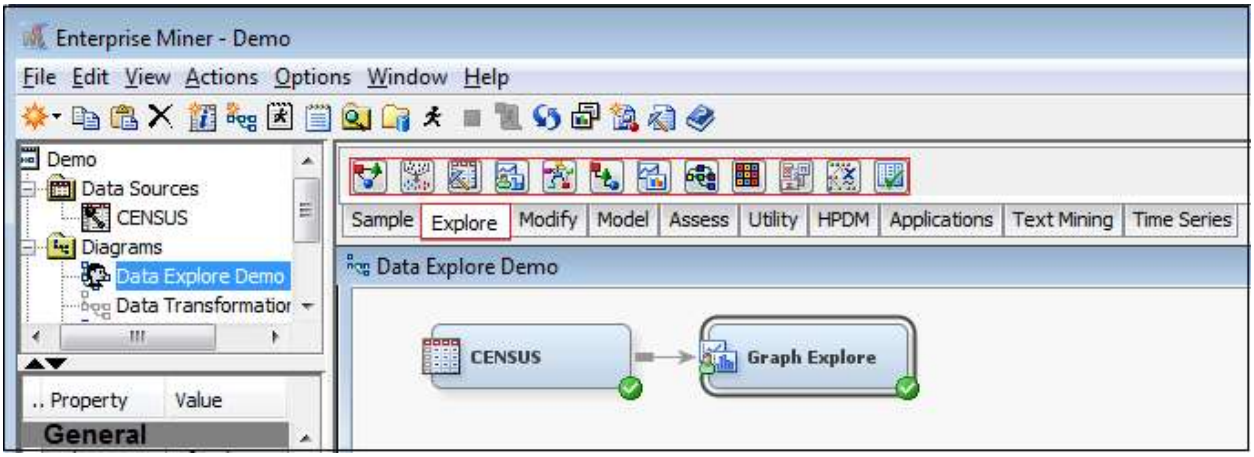


Figure 10. Univariate Analyses Using *Graph Explore* Node in SAS ENTERPRISE MINER®

5. VARIABLE SELECTION (a.k.a.: VARIABLE REDUCTION, FEATURE SELECTION)

When there are over hundreds or even thousands of variables after including various internal and external data sources, the variable selection process becomes critical and challenging. Besides noise, redundancy and irrelevancy are the two keys to reduce variables in the variable selection process. Redundancy means the variable doesn't provide any additional new information that other variables have already provided. Irrelevancy means that the variable doesn't provide any useful information about the target. See some common variable selection techniques below:

1) Correlation Analysis: Identify variables which are correlated to each other to avoid multicollinearity to build a more stable model

2) Multivariate Analyses: Cluster Analysis, Principle Component Analysis, and Factor Analysis.

Some common SAS procedures as follows:

The CORR Procedure, the VARCLUS Procedure, and the FACTOR Procedure. PRINCPLE procedure

3) Stepwise Selection Procedure: *Stepwise selection* is a method that allows moves in either direction, dropping or adding variables at the various steps.

Backward stepwise selection *(Backward Elimination)* starts with all the predictors to remove the least significant variable, and then potentially add back variables if they later appear to be significant. The process is one of alternation between choosing the least significant variable to drop and then re-

considering all dropped variables (except the most recently dropped) for re-introduction into the model. This means that two separate significance levels must be chosen for deletion from the model and for adding to the model. The second significance must be more stringent than the first.

Forward stepwise selection is also a possibility, though not as common. In the forward approach, variables once entered may be dropped if they are no longer significant as other variables are added.

Stepwise Regression is a combination of backward elimination and forward selection. This addresses the situation where variables are added or removed early in the process and we want to change our mind about them later. At each stage a variable may be added or removed and there are several variations on exactly how this is done.

The simplified SAS code below shows how a stepwise logistic regression procedure to select variables for a logistic regression model (binary target variable) can be used in Base SAS®/SAS Enterprise Guide®.

```
proc logistic data = datasample;
   model litigation_flag (event = '1')= var1 var2 var3
   /selection=stepwise  slentry=0.05  slstay=0.06;
   output out=datapred1 p=phat lower=lcl upper=ucl;
run;
```

In SAS Enterprise Miner®, the three most commonly used variable selection nodes in SAS Enterprise Minder® are introduced below. The ***Random Forest*** **Node** also could be used for variable selection purpose but are less commonly used.

The *Variable Selection* Node can be used for selecting variables. This procedure provides a tool to reduce the number of input variables using R-square and Chi-square selection criteria. The procedure identifies input variables which are useful for predicting the target variable and ranks the importance of these variables. The diagram below (Figure 11) contains an example.

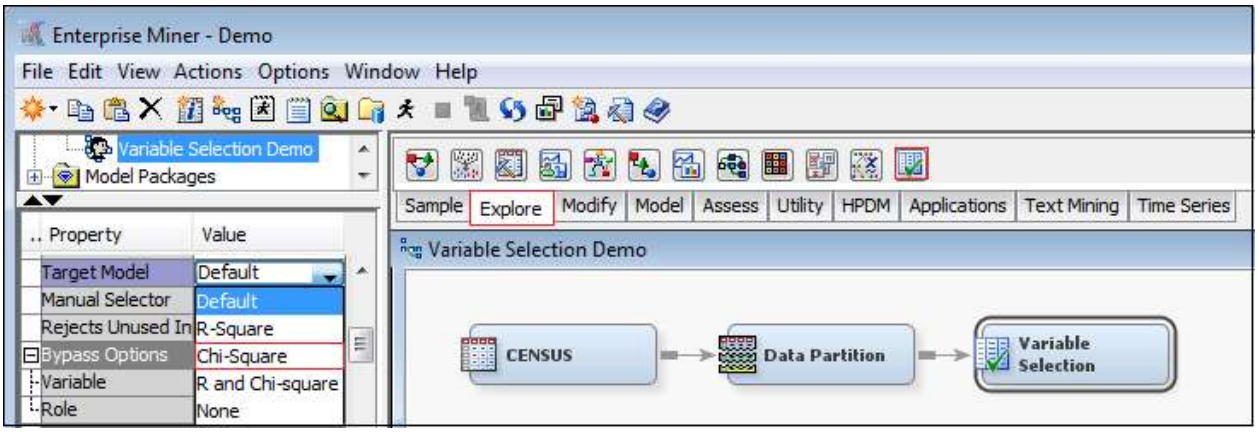


Figure 11. A Variable Selection Example Using *Variable Selection* Node in SAS Enterprise Miner®

The *Regression* Node is used for selecting variables involves using to specify a model selection method. If ***Backward*** is selected, training begins with all candidate effects in the model and removes effects until the Stay significance level or the stop criterion is met. If ***Forward*** is selected, training begins with no candidate effects in the model and adds effects until the Entry significance level or the stop criterion is met. If ***Stepwise*** is selected, training begins as in the Forward model but may remove effects already in the model. This continues until the Stay significance level or the stop criterion is met. If ***None*** is selected, all inputs are used to fit the model. The diagram below (Figure 12) contains an example.

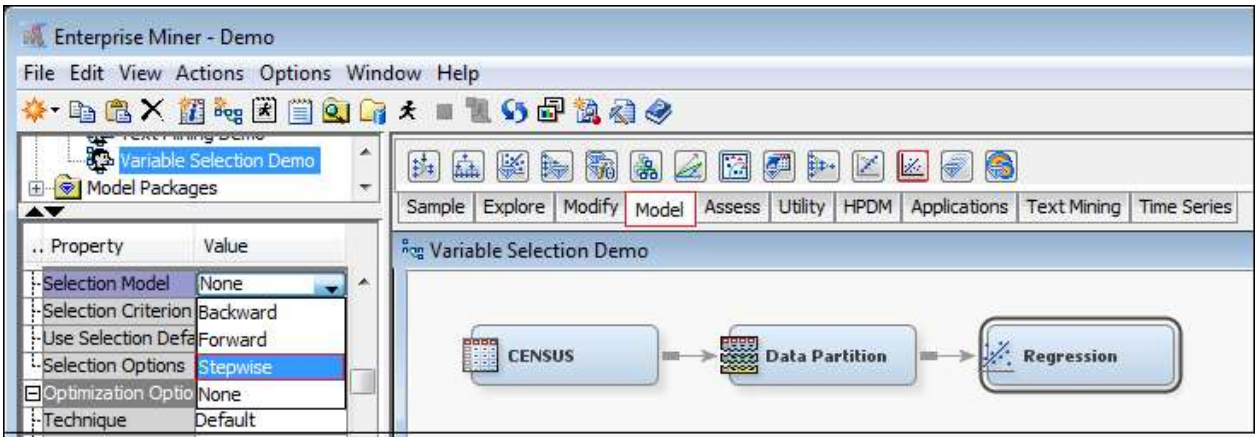


Figure 12. A Variable Selection Example Using *Regression* Node in SAS Enterprise Miner®

The *Decision Tree* Node can be used for selecting variables. This procedure provides a tool to reduce the number of input variables by specifying whether variable selection should be performed based on importance values. If this is set to ***Yes***, all variables that have an importance value greater than or equal to 0.05 will be set to Input. All other variables will be set to Rejected. The diagram below (Figure 13) contains an example.

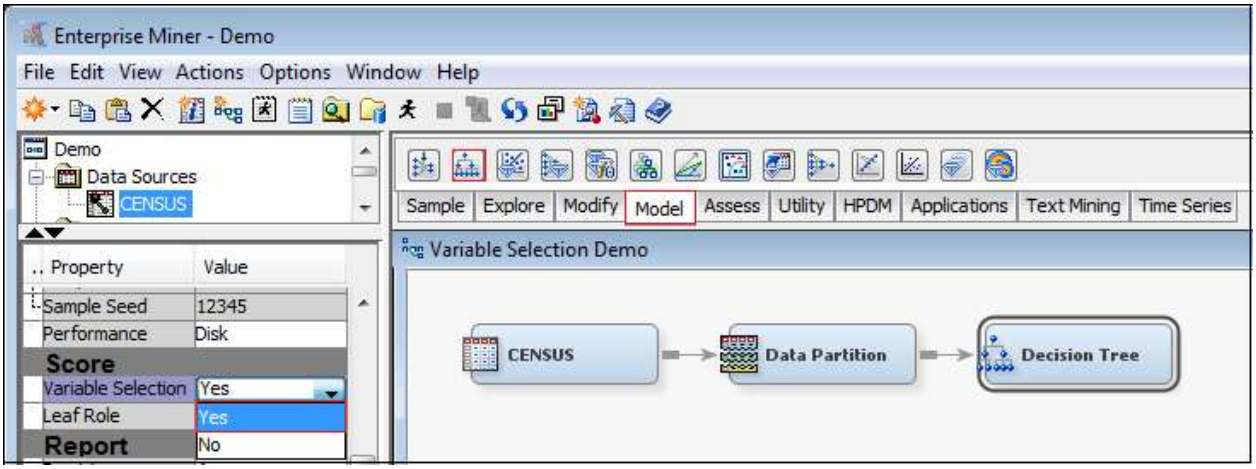


Figure 13. A Variable Selection Example Using *Decision Tree* Node in SAS Enterprise Miner®

6. MODEL BUILDING (a.k.a.: MODEL FITTING)

An insufficient model might systematically miss the signal(s), which could lead to high bias and underfitting. Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trends and patterns of the data.

Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. The overly complicated model might perform well on a specific data set but mistakes the random noise in the data set for signal; such a model may lead to misleading results if it is used to predict on other data sets.

There are usually many iterations to fit models until the final model which is based on both desired statistics (relatively simple, high accuracy, and high stability) and business application. The final model includes the target variable, independent variables, and multivariate equations with weights and coefficients for the variables used.

The Generalized Linear Modeling (GLM) technique has been popular in the property and casualty insurance industry for building statistical models.

The Generalized Linear Modeling (GLM) technique has been popular in the P&C insurance industry for building statistical models. There are usually many iterations to fit models until the final model, which is based on both desired statistics and business application. Below is a simplified SAS code of a logistic regression fit using PROC GENMOD (GLM) in Base SAS®/SAS Enterprise Guide®:

```
proc genmod data=lib.sample;
  class var1 var2 var3 var4;
  model litigation_flag = var1 var2 var3 var4 var5
  /dist = bin
  link=logit lrci;
  output out=lib.sample p=pred;
run;
```

The same GLM logistic regression procedure can be done using ***Regression*** node with specifying model selection as ***GLM*** in SAS Enterprise Miner®. The diagram below (Figure 14) contains an example.

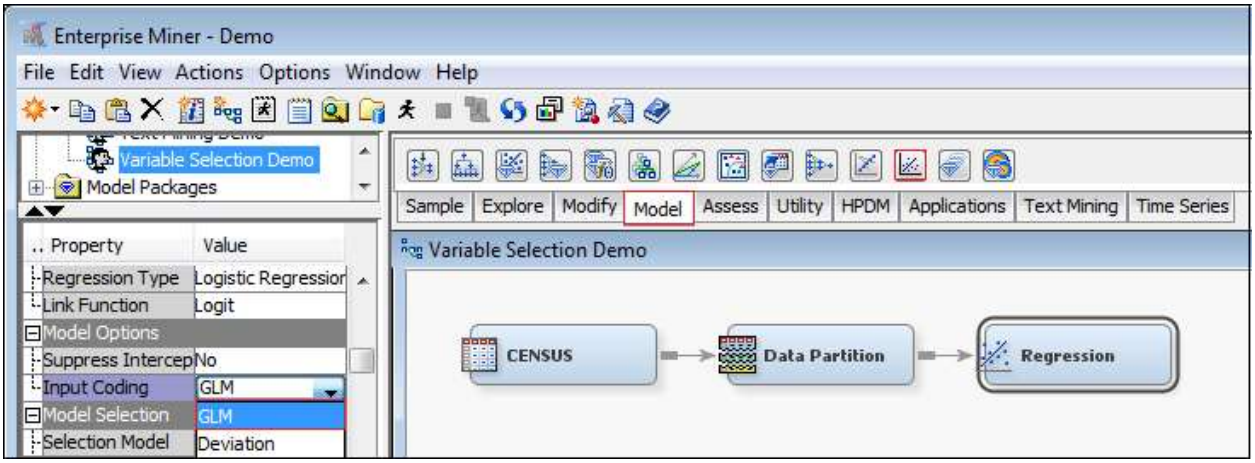


Figure 14. A GLM Logistic Regression Example Using *Regression* Node in SAS Enterprise Miner®

Interaction and correlation usually should be examined before finalizing the models. Other model building/fitting methodologies could be utilized to build models in SAS Enterprise Miner® including the following three types of models (The descriptions below are attributable to SAS Product Documentation):

Decision Tree Model is a predictive modeling approach which maps observations about an item to conclusions about the item's target value. A decision tree divides data into groups by applying a series of simple rules. The rules are organized hierarchically in a tree-like structure with nodes connected by lines. The first rule at the top of the tree is called the *root node*. Each rule assigns an observation to a group based on the value of one input. One rule is applied after another, resulting in a hierarchy of groups. The hierarchy is called a tree, and each group is called a node. The original group contains the entire data set and is called the root node of the tree. A node with all its successors forms a branch of the node that created it. The final nodes are called leaves. For each leaf, a decision is made and applied to all observations in the leaf. The paths from root to leaf represent classification rules.

Random Forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

Gradient Boosting Model uses a partitioning algorithm to search for an optimal partition of the data for a single target variable. Gradient boosting is an approach that resamples the analysis data several times to generate results that form a weighted average of the resampled data set. Tree boosting creates a series of decision trees that form a single predictive model. Like decision trees, boosting makes no assumptions about the distribution of the data. Boosting is less prone to overfit the data than a single decision tree. If a decision tree fits the data fairly well, then boosting often improves the fit.

Neural Network Model: Organic neural networks are composed of billions of interconnected neurons that send and receive signals to and from one another. Artificial neural networks are a class of flexible nonlinear models used for supervised prediction problems. The most widely used type of neural network in data analysis is the multilayer perceptron (MLP). MLP models were originally inspired by neurophysiology and the interconnections between neurons, and they are often represented by a network diagram instead of an equation. The basic building blocks of multilayer perceptrons are called hidden units. Hidden units are modeled after the neuron. Each hidden unit receives a linear combination of input variables. The coefficients are called the (synaptic) weights. An activation function transforms the linear combinations and then outputs them to another unit that can then use them as inputs.

Support Vector Machine (SVM) is a supervised machine-learning method that is used to perform classification and regression analysis. The standard SVM model solves binary classification problems that produce non-probability output (only sign +1/-1) by constructing a set of hyperplanes that maximize the margin between two classes. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification.

Model Ensemble creates new models by combining the posterior probabilities (for class targets) or the predicted values (for interval targets) from multiple predecessor models. There are three methods: Average, Maximum, and Voting.

Figure 15 shows a simplified modeling procedure with most popular models in the SAS Enterprise Miner®.

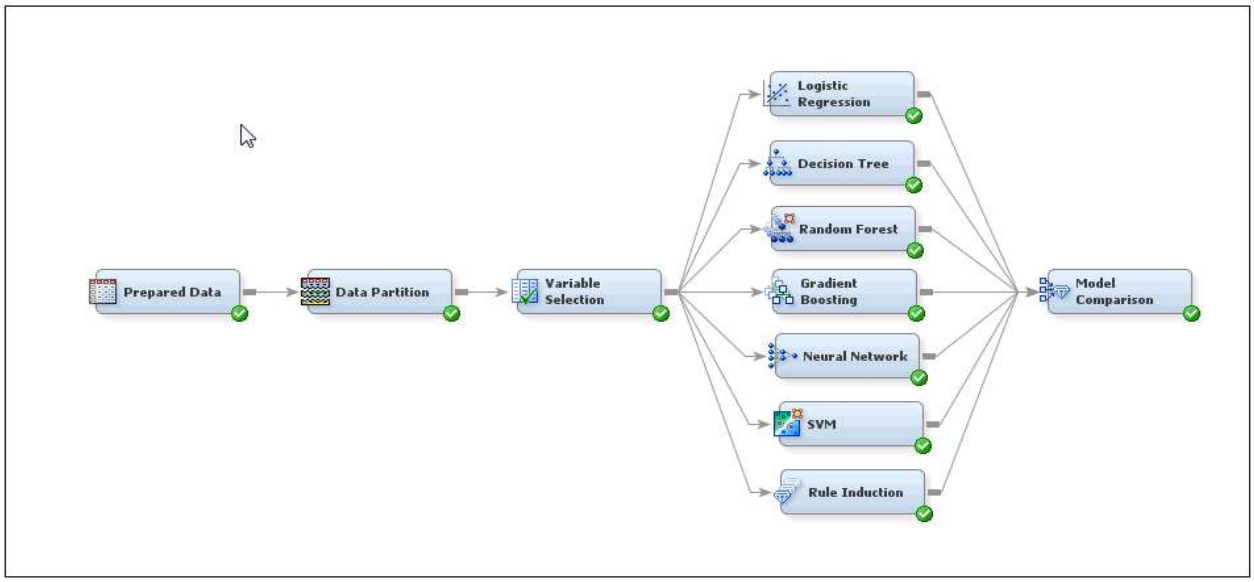


Figure 15. A Simplified Example of Model Building Procedure with Seven Common Models

7. MODEL VALIDATION

Model validation is a process to apply the candidate models on the validation data set to select a best model with a good balance of model accuracy (including precision and recall ratios, etc.) and stability. Common model validation methods include Gini Ratios, Lift Charts, Confusion Matrices, Receiver Operating Characteristic (ROC), Gain Charts, Bootstrap Sampling, and Cross Validation, etc. to compare actual values (results) versus predicted values from the model. Bootstrap Sampling and Cross Validation methods are especially useful when data volume is concerned.

Cross Validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set and it can detect overfitting problem. Cross Validation data splitting is introduced through the simplified example (Figure 16) below. Four cross fold subset data sets are created for validating stability of parameter estimates and measuring lift. The diagram shows one of the cross folds with 60% random training data, the other three cross folds would be created by taking other combinations of the Cross Validation data sets. In Cross Validation, a model is fitted using 60% random training data (Cross Fold 1), then the model is used to score the rest 20% of the data. The process is repeated three times for the other three cross fold subset data sets.

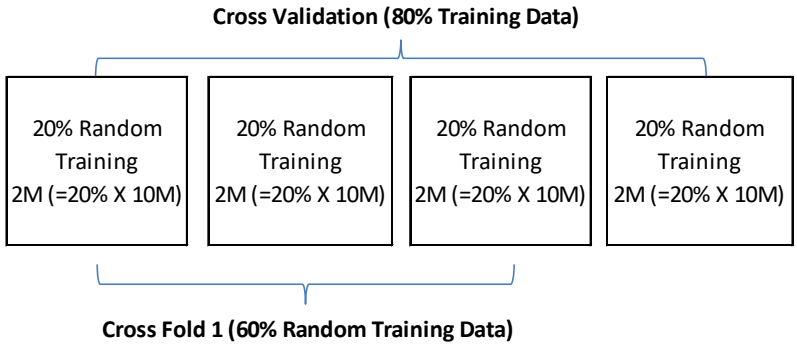


Figure 16. A Diagram of Creating Simplified Cross Four Fold Subset Data Sets for Cross Validation

In SAS Enterprise Miner®, run the model fitting (Logistics Regression or Decision Tree, etc.) on each of the cross fold subset data sets to get parameter estimates. Then examine the four sets of parameter estimates side by side to see if they are stable. The following diagram (Figure 17) shows a Cross Validation process in SAS Enterprise Miner®. A macro could be created and utilized to run the same model fitting process on four cross fold subset data sets in SAS Enterprise Guide®.

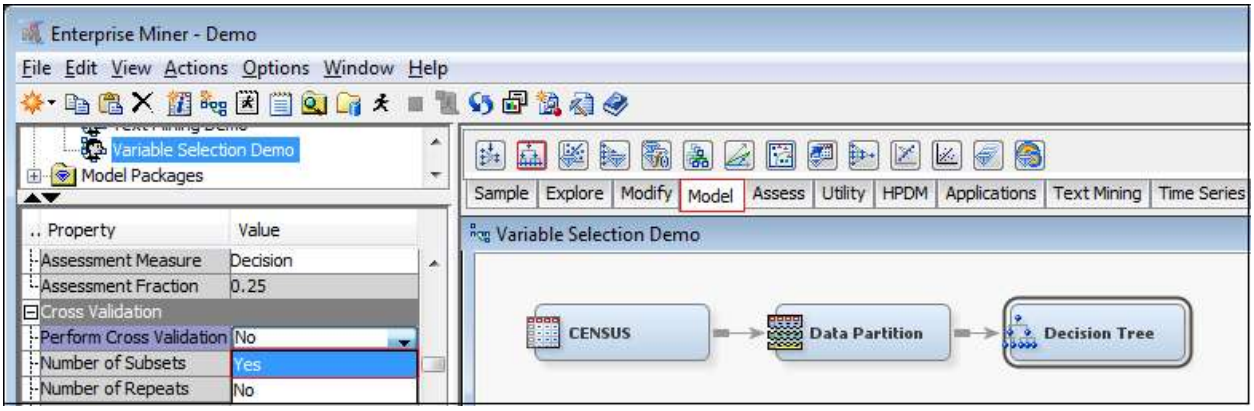


Figure 17. A Cross Validation Simplified Example Using *Decision Tree* node in SAS Enterprise Miner®

The following common fit statistics are reviewed for model validation: Akaike's Information Criterion, Average Squared Error, Average Error Function, and Misclassification Rate, etc.

8. MODEL TESTING

The overall model performance on validation data could be overstated because the validation data has been used to select the best model. Therefore, model testing is necessarily performed to further evaluate the model performance and provide a final unbiased assessment of model performance. Model testing methods are similar as model validation but using holdout testing data and/or new data.

For litigation propensity model, we can also test the model by industries, benefit states, body parts, clients, etc. to verify and understand the model performance.

The following diagram (Figure 18) shows a predictive modeling process with major stages starting from prepared data to data partition, variable selection, building logistic regression model, and testing (scoring) new data in SAS Enterprise Miner®.

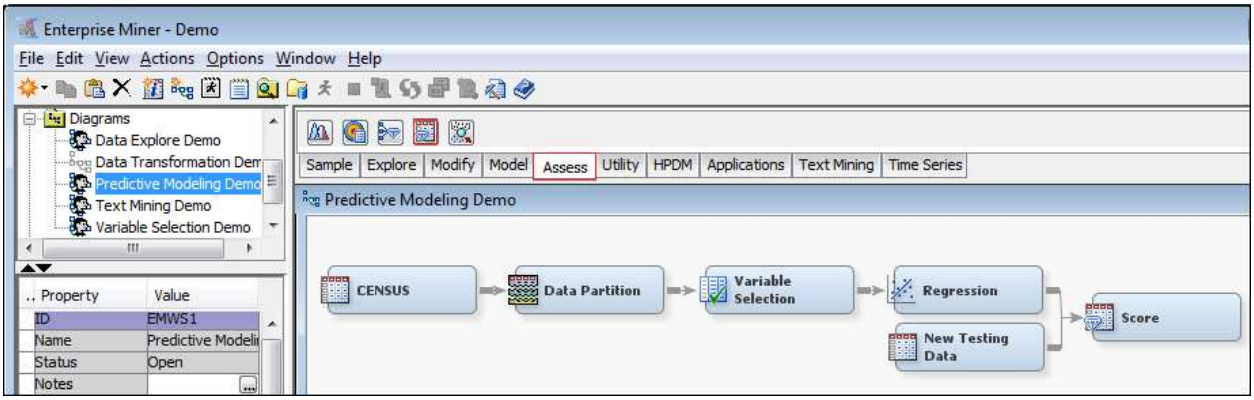


Figure 18. A Simplified Predictive Modeling Flow Chart in SAS Enterprise Miner®

9. MODEL IMPLEMENTATION AND MONITORING

The last very important stage of any predictive modeling project is the model implementation which is the stage to turn all the modeling work into action to achieve the business goal and solve the business problem. Usually before model implementation, a model pilot would be helpful to get some sense of the model performance and prepare for the model implementation appropriately.

If there is an existing model, we would also like to conduct a model champion challenge to understand the benefit of implementing the new model over the old one. There are numeric ways to implement the models which largely depends on what type of models and the collaboration with IT support at the organization.

The most important aspect of any predictive modeling project is the result, and whether the result solves the business problem, adds value to the organization, or fulfills the business goal(s). Therefore, model performance monitoring and results evaluation should be conducted periodically to see if the model is still performing well to generate reasonable results. If not, then rebuilding the model by including the data from most recent years and/or more data sources should be explored and considered.

CONCLUSION

The primary goal of this paper is to provide an overview of claim analytics and introduce a general modeling process, as part of industry standard practice, through a litigation prediction case study in SAS Enterprise Guide® and SAS Enterprise Miner®. This process could be tweaked and applied to different claim handling processes and different business goals for different insurance companies and third party administrators. The process could also be used in other statistical software, like R and Python.

Claim analytics, just like analytics in any other industry, has been developing and evolving along with new technologies and new data sources to solve more claim business problems, further improve claim handling efficiency to reduce claim cost, and help claimants recover faster and return to work earlier.

REFERENCES

Census Data Source (http://www.census.gov/)

SAS Product Documentation (http://support.sas.com/documentation/)

Towers Watson Survey

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Mei Najim, Founder and Lead Data Scientist, Advanced Analytics Consulting Services, LLC.

E-mail: mei_najim@aacsus.com

LinkedIn: https://www.linkedin.com/in/meinajim/