

## Reevaluating Policy and Claims Analytics: a Case of Non-Fleet Customers In Automobile Insurance Industry

Kittipong Trongsaewad and Jongsawas Chongwatpol

NIDA Business School, National Institute of Development Administration, Bangkok, Thailand

### ABSTRACT

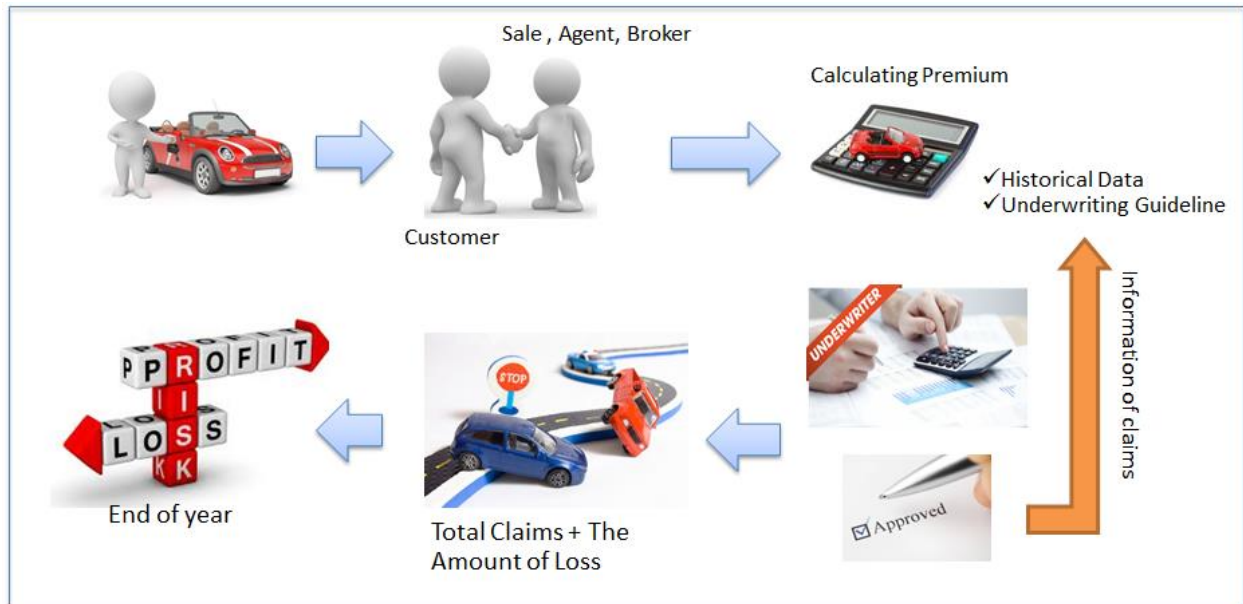
Analyzing automobile policies and claims is an ongoing area of interest to the insurance industry. Although there have been many data mining projects in insurance sector over the past decade, the following questions - *How can insurance firms retain their best customers? Will this damaged car be covered and get claim payment? How much of loss of claims associated with this policy will be?* – do remain as common. This study applies data mining techniques using SAS® Enterprise Miner™ to enhance insurance policies and claims. The main focus is on assessing how corporate fleet customers' policy characteristic and claim behavior are different from that of non-fleet customers. We believe that implementing advanced analytics can help create better planning for policy and claim management strategy.

### 1. INTRODUCTION

XYZ Company (name disguised) is one of the top ten leading non-life Insurance companies in Thailand, is a stable company and listed in SET with the registered capital of THB 590 Million, and is professionalism in insurance and the non-life Insurance company of the Thai people with professional ethics, transparency and good corporate governance. The company's main specialization can be categorized as motor and non-motor insurance. Figure 1 presents the general process flow on how a policy is issued.

As presented in Figure 1, the decision to issue or decline a policy are made in step 3, where insurers' background and claim history are checked. However, such assessment is not applicable for new customers. Additionally, because the profit and loss of motor insurance business can be recognized after the policy is expired. In fact, it takes the whole policy year to receive the result of business profitability which conduces to the strategy development to achieve the target of the company. Meanwhile, the company would provide the policy coverage and does not know that policy will return the amount of loss until the claims have been incurred.

The following business questions - *How can insurance firms retain their best customers? Will this damaged car be covered and get claim payment? How much of loss of claims associated with this policy will be?* – do remain as common. This study applies data mining techniques using SAS® Enterprise Miner™ to enhance insurance policies and claims. The main focus is on assessing how corporate fleet customers' policy characteristic and claim behavior are different from that of non-fleet commercial customers. Thus, implementing advanced analytics help create better planning for policy and claim management strategy.



**Figure 1: General Process Flow of Issued Automobile Policy**

## 2. DATA PREPARATION

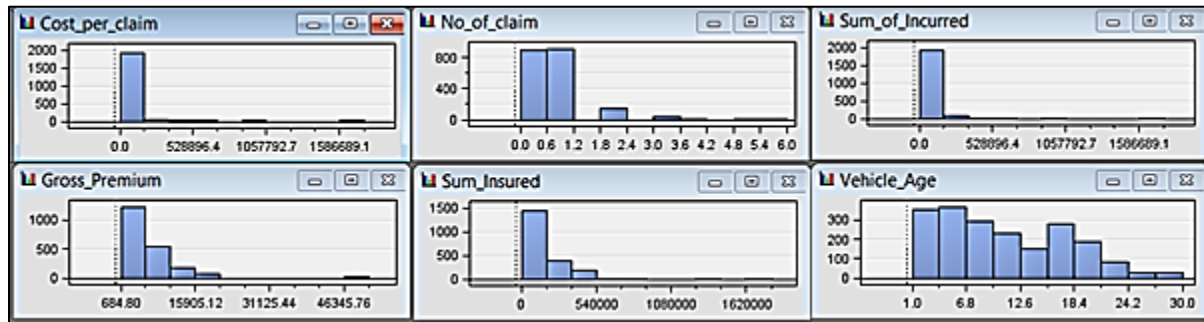
We follow the five steps of the SEMMA methodology– Sample, Explore, Modify, Model, and Assess. We first explore all insurance related and non-insurance related variables to get a sense of both policy and claim datasets. Table 1 presents the summary of both dependent and independent variables used in this study. Below are the key findings in the data preparation process:

- The data source used to perform data mining analysis in this study is categorized into two groups: Fleet (Dataset Name: PolicynClaims2012Fleet) and NonFleet (Dataset Name: PolicynClaims2012nonFleet).
- The dataset contains the total of 25 variables related to both insurance and non-insurance related factors.
- Sum\_of\_Incurred (the dependent variable) refers to the total of claims incurred for all policies
- New customers can be determined based on the original effective date of policy coverage; thus 1 = new customers and 0 = renewed customers.
- There are three different types of policy coverage where “CO” refers to Comprehensive Motor Insurance, “TP” refers to Third Party Only Motor Insurance, and “TF” refers to Theft and Fire Motor Insurance.
- Based on the policy coverage, the policy type “CO” must not have gross premium lower than 7,000B.
- Any vehicles with the age over 30 years are removed from this analysis.

**Table 1: Data Description**

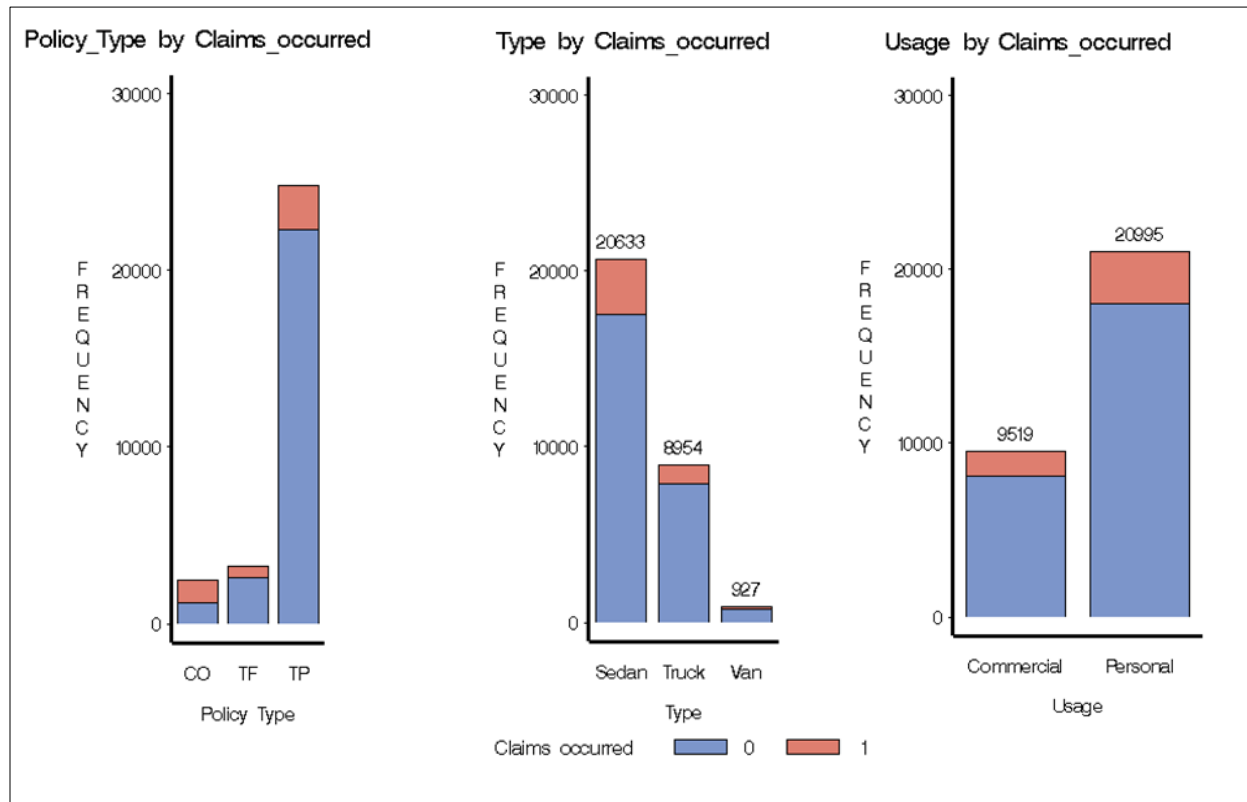
Variable Type	Variable Name	Level	Description
Dependent (Target)	Claims_occured	Binary	If the policy has claim, the value is 1
	Sum_of_Incurred	Interval	Total of claims occurred between effective date and expire date of the policy
Independent	Capacity	Nominal	Engine size
	Channel	Nominal	Sale channel
	Cost_per_claim	Interval	
	Deductible	Interval	The amount of payment that the customer must pay before claim which specify as the policy conditions
	Gross_Premium	Interval	Total premium of the policy
	newp	Binary	Type of the policy, new or renew
	No_of_claim	Interval	The number of claim occurred in the policy coverage year
	Policy_No	Nominal	ID of the policy
	Policy_Type	Nominal	The coverage type of the policy
	ProvinceCode	Nominal	Code of province
	Sum_Insured	Interval	The maximum amount of liability
	Type	Nominal	Body type of the car
	Usage	Nominal	Character of usage
	Vehicle_Age	Interval	Age of car
	Vehicle_Make	Nominal	Brand of car
	Original Effective Date	Nominal	The first effective date that record before endorsement
	Effective Date	Nominal	
	Expire Date	Nominal	
	coverdate	Interval	The number of coverage days
	Fleet	Nominal	The type of recording policy
	CNTTYP	Nominal	The type of the policy such as PMX=Voluntary, CTP=Compulsory
	Province	Nominal	The province that the car is registered in Thailand
	Claim_No	Nominal	ID of the claim

Figure 2 presents an example of distribution of interval variables such as Cost\_per\_claim, Gross\_Premium, No\_of\_claim, Sum\_Insured, Sum\_of\_Incurred and Vehicle\_Age. We can see that all variables are right skewed and variable transformation is evaluated later on whether the predicted results get better.



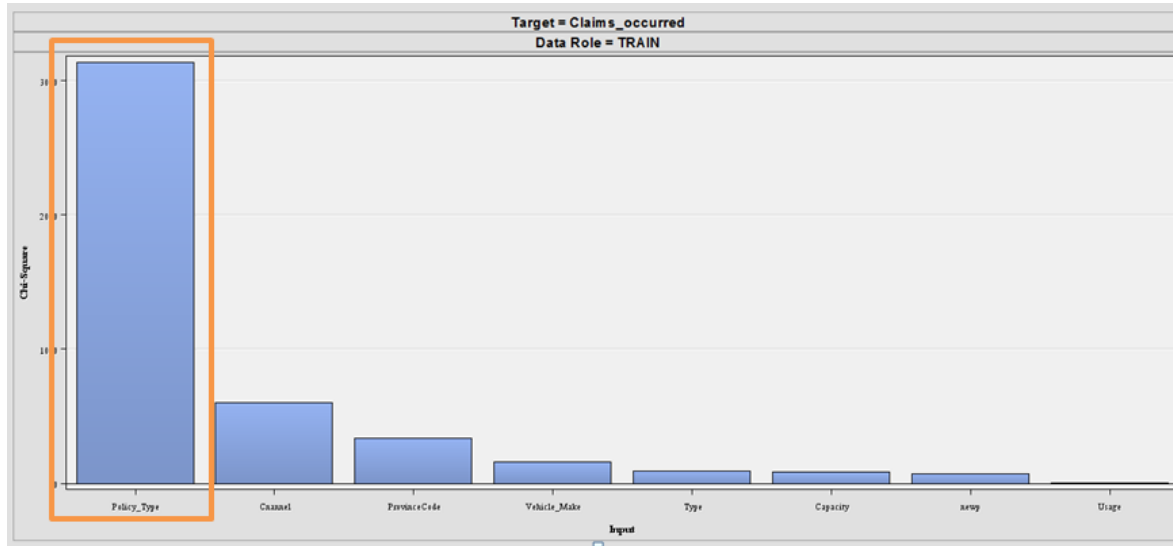
**Figure 2: The Histogram of Independent Variables**

Our next task is to explore the characteristics of claims occurred in 2012 by the policy type, vehicle type, and usage type. As presented in Figure 3, claims occur mostly with the “Third Party Only Motor Insurance or TP” policy. When looking at the type of vehicles, van and truck have fewer claims than sedan; meanwhile claims occur mostly on the personal vehicles.



**Figure 3: Number of Claims by Policy Type, Vehicle Type, and Usage**

Based on the Chi Square plot in Figure 4, Policy Type is strongly associated with the target variable “Claims\_Occurred”, followed by Channel, Province Code, and Vehicle type.



**Figure 4: Chi Square Plot for Target Variable “Claims\_Occurred”**

After data exploration, any inconsistencies, errors, or extreme values in the dataset are treated appropriately. Some variables are transformed for better model development. We then categorize our analysis into two different groups based on insurers’ characteristic. The first group is for corporate insurers, who own a fleet of vehicles that are primarily used for business purposes. The second group is a not-fleet group, which mainly focuses on retail customers. Currently, the underwrite processes for both groups are similarly based on their background and claim history. Understanding their characteristics of each type of customers can help in issuing new policy more effectively. Figure 5 presents the research framework of this study.

### 3. MODELING

As presented in Figure 5, the final dataset is partitioned into training and testing dataset. For the first model, three popular data mining techniques including decision tree, logistic regression, and neural networks are used to predict whether claims have been incurred during the policy year or not. Ensemble model is also considered to ensure the best model for claim prediction.

The complicity of the model is controlled by fit statistics calculated on the testing dataset. We use three different criteria to select the best model on the testing dataset. These criteria include false negative, prediction accuracy, and misclassification rate. False negative (Target = 1 and Outcome = 0) represents the case of an error in the model prediction where model results indicate that diabetes occurrence is not present, when in reality, there is an incident. The false negative value should be as low as possible. The proportion of cases misclassified is very common in the predictive modeling. However, the observed misclassification rate should be also relatively low for model justification. Lastly, prediction accuracy is evaluated among the three models on the testing dataset. The higher the prediction accuracy rate, the better the model to be selected.

The result of the first model is then further analyzed to determine which claims are most likely to generate cost savings. In the second model, the main goal is to predict “loss of claims”. Consequently, the overall profit of the insurance company can be estimated. Additionally, we then assess both policy and claim behavior of both corporate fleet and no-fleet customers.

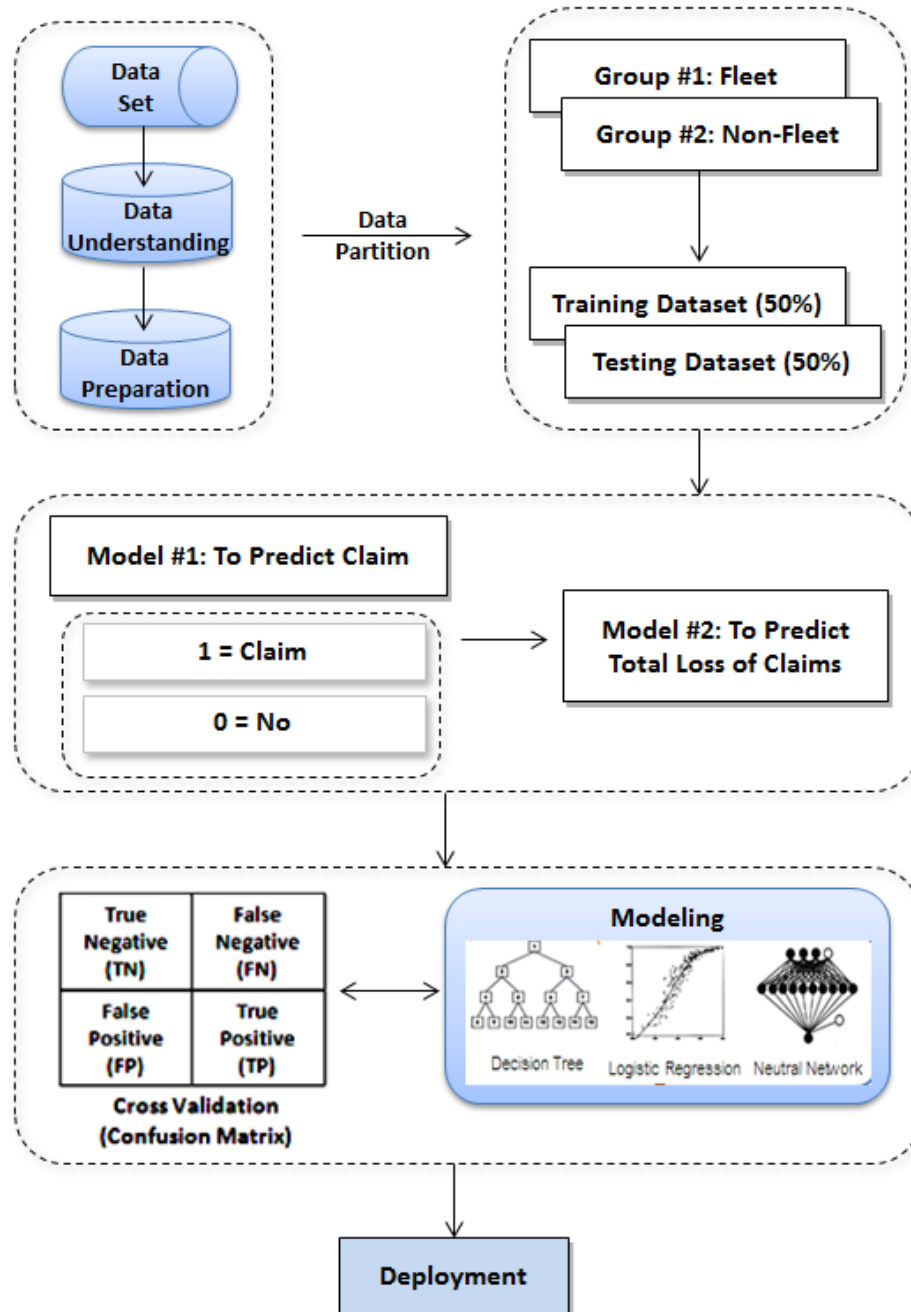
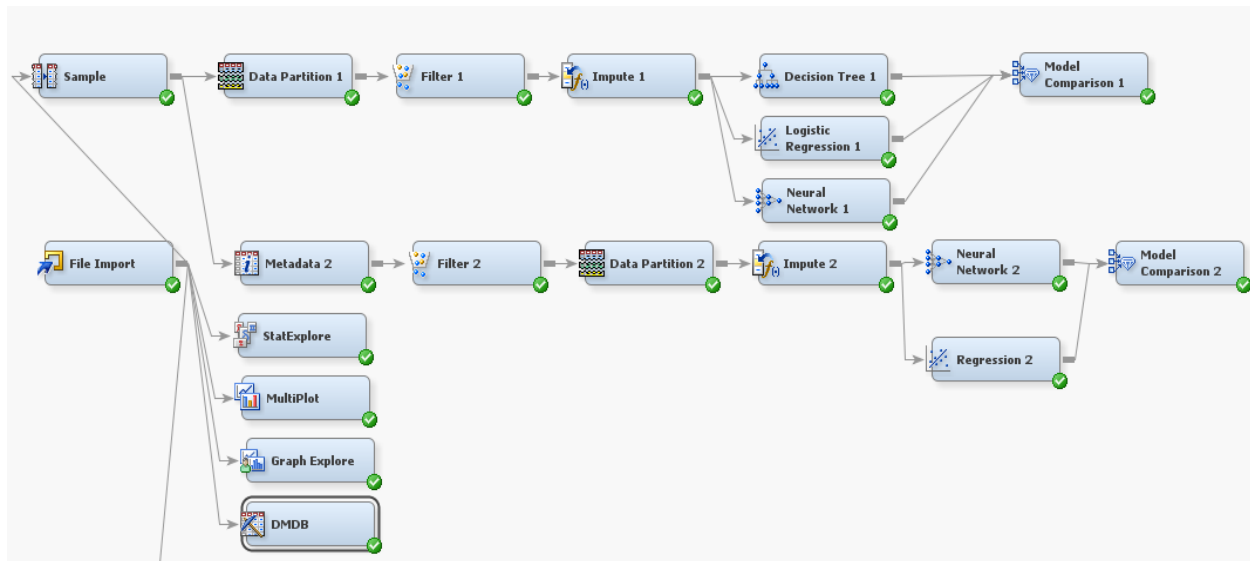


Figure 5: Research Framework

In this study, we also decided to stratify our samples so that the number of claims (Target variable “Claims\_Occurred” = 1) and non-claims (Target variable “Claims\_Occurred” = 0). With a 50% adjusting for oversampling, the prediction results on the balanced dataset are more appropriated since the original dataset contains approximately only 15% of claims occurred (Clams\_Occurred = 1). Figure 6 presents the flow diagram of this study.



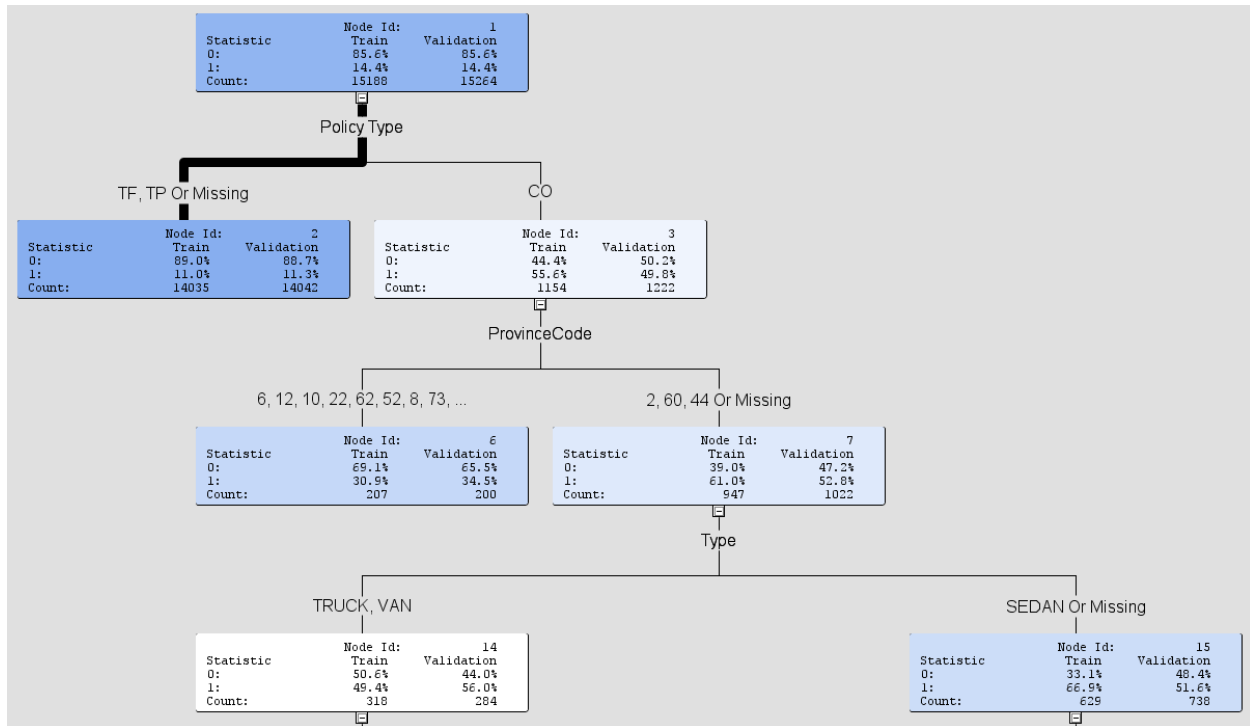
**Figure 6: SAS® Enterprise Miner™ – Flow Diagram**

#### 4. RESULTS AND DISCUSSION

To push advanced analytics to aid in predicting claims, the model to predict the probability of the incidence of claims is built. The results of decision trees are presented are presented in Figure 7. The following examples illustrate rule-based prediction algorithms:

- IF the Policy Type is [TF or TP], the probability of the incidence of claims is 11.3%
- IF the Policy Type is [CO] and the province code is [6, 12, ...] THEN the probability of the incidence of claims is 34.5%
- IF the Policy Type is [CO] and the province code is [2, 60, ...] and type is [Sedan] and vehicle Age is [>4.5] THEN the probability of the incidence of claims is 49%

Based on the Decision Tree Model in Figure 7, Policy\_Type, ProvinceCode, Capacity, Vehicle\_Make, and Channel are considered “the variable importance”, which are used for splitting the tree nodes in the model. However, based on the model comparison in Figure 8, Neural Network produces the best results with the overall misclassification rate of 13.76%, followed by Decision Tree (13.89%) and Logistic Regression (14.01%)



#### Variable Importance

Obs	NAME	LABEL	NRULES	IMPORTANCE	VIMPORTANCE	RATIO
1	Policy_Type	Policy Type	2	1.0000	1.0000	1.0000
2	ProvinceCode	ProvinceCode	1	0.2543	0.1597	0.6277
3	Capacity	Capacity	1	0.1771	0.0000	0.0000
4	Vehicle_Make	Vehicle Make	1	0.1242	0.0000	0.0000
5	Cchannel	Cchannel	1	0.1149	0.2135	1.8579

Figure 7: An Example of Decision Tree Models

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Total Degrees of Freedom	Train: Degrees of Freedom for Error
Y	Neural3	Neural3	Neural Net...	Claims_oc...	Claims occ...	0.137602	14863.85	14484.85
	Tree2	Tree2	Decision Tr...	Claims_oc...	Claims occ...	0.138956	14863.85	.
	Reg3	Reg3	Logistic Re...	Claims_oc...	Claims occ...	0.140142	14863.85	14768.85

Figure 8: Model Comparison



For the groups with high probability of the incidence of claims, we can further analyze the total loss of claims associated with the policy. By building a stepwise regression model, the results show that Capacity, Channel, Policy\_type, Sum\_Insured, Vehicle\_Age, and newp are the key important factors to explain and predict the total loss of claims. Figure 8 presents the summary of stepwise selection as indicated.

Summary of Stepwise Selection					
Step	Effect Entered	DF	Number In	F Value	Pr > F
1	Policy_Type	2	1	103.45	<.0001
2	Capacity	19	2	4.54	<.0001
3	Sum_Insured	1	3	14.26	0.0002
4	Cchannel	4	4	4.62	0.0010
5	Vehicle_Age	1	5	12.03	0.0005
6	newp	1	6	4.31	0.0380

The selected model is the model trained in the last step (Step 6). It consists of the following effects:

Intercept Capacity Cchannel Policy\_Type Sum\_Insured Vehicle\_Age newp

**Figure 8: Summary of Stepwise Selection**

The result of the first model is then further analyzed to determine which claims are most likely to generate cost savings. In the second model, the main goal is to predict “loss of claims”. Consequently, the overall profit of the insurance company can be estimated. Additionally, we then assess both policy and claim behavior of both corporate fleet and no-fleet customers. We believe that implementing advanced analytics can help create better planning for policy and claim management strategy.

## 5. CONCLUSION

This data mining project helps develop a program to aid in the underwriting process. The scoring model gives a probability of a given insurance applicant defaulting on claims and the total loss of claims. The threshold can be selected such that all applicants whose probability of defaults is in excess of the threshold level (80%, for instance) will be recommend for rejection, a closer attention, or higher deductible options. Some other key findings of this study are as follows:

- When gross premium increases, the total loss of claims decreases.
- When the maximum amount of liability increases, the total loss of claims decreases.
- The non-Japanese vehicles have higher chance to report claims than the Japanese vehicles.
- Sedan vehicles have the lowest risk of loss of claims; meanwhile truck has the highest risk of loss of claims
- For the market channel, direct sale has the lowest risk of loss of claims compared to Agent, Broker. Motor Partner and Kbank sales.

For the deployment of this study, we apply the xml code of the first and the second model in the program which is used for transferring the e-application, batch file in text or excel formats that partners have sent to underwrite and issue the policy to the database. This program will support the preliminary information for underwriter to make a decision to accept the application.

Lastly, one of the lesson learns from this study is on the data preparation process. There are several issues regarding the quality of data, missing several key insurance-related data, data duplication, and data inconsistency while conducting this research. Thus, in addition to the implementation of the models to predict claims and the total loss of claims, the company should (1) define what variable should be collected and included in the analysis, (2) set those variables as compulsory fields to force users to entry the data, (3) improve the data entry system to validate the data completeness, and (4) specify the standard of data entry, especially for internal users.

## CONTACT INFORMATION

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

Name: Kittipong Trongkawad,  
Enterprise: NIDA Business School, National Institute of Development Administration  
Address: 300/778 Supalai city resort ratchada-huaykwang, Prachauthid Road, Huaykwang ,  
Bangkok, 10310 Thailand  
Email: ktpong\_ice@hotmail.com

Name: Jongsawas Chongwatpol, Ph.D.  
Enterprise: NIDA Business School, National Institute of Development Administration  
Address: 118 Seri Thai Road, Bangkapi, Bangkok, 10240 Thailand  
Email: jongsawas.c@ics.nida.ac.th, jong\_tn@hotmail.com

Kittipong Trongkawad is an MBA Student at National Institute of Development Administration. He received his BE in industrial engineering from Kasetsart University, Bangkok, Thailand. He works at the one of the top ten insurance companies in Thailand. He is currently working as a process development officer. His main task is to advise and coordinate with associates to improve the efficiency of the operating department, which starts up from Front line to End line. Additionally, he also provides a support to the motor insurance processing department by employing both statistical, analytical and improvement tool to create the long-term strategy to achieve the assigned action plan. He is also interested in the field of data mining and business intelligences.

Jongsawas Chongwatpol is a lecturer in NIDA Business School at National Institute of Development Administration. He received his BE in industrial engineering from Thammasat University, Bangkok, Thailand, and two MS degrees (in risk control management and management technology) from University of Wisconsin - Stout, and PhD in management science and information systems from Oklahoma State University. His research has recently been published in major journals such as Decision Support Systems, Decision Sciences, European Journal of Operational Research, and Journal of Business Ethics. His major research interests include decision support systems, RFID, manufacturing management, data mining, and supply chain management.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.