Lowering social cost of car accidents by predicting high-risk drivers

Vannessa Peng
Davin Tsai
Shu-Min Yeh

Why we do this?

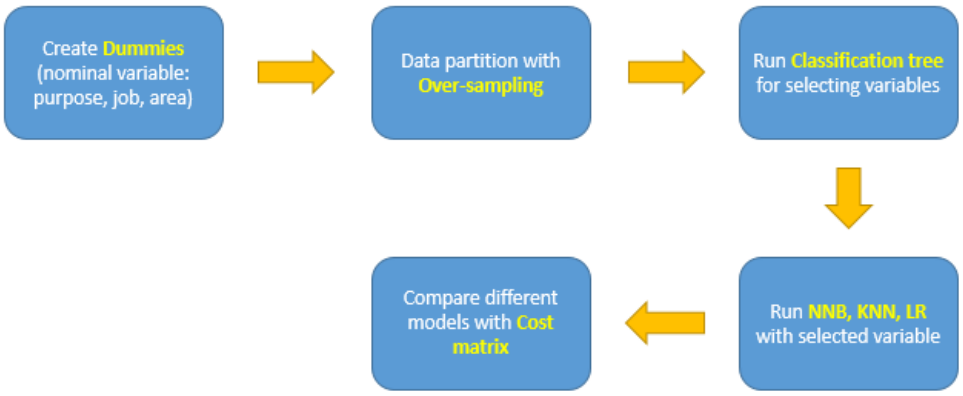
Traffic accident happened every day. In order to decrease the number of traffic accident and the losses caused by the traffic accident. Our government have a lot of policy, but these policy are focus on all the citizen. But now we think the policy have to focus on the certain group, the group who have high rate in traffic accident, to let the policy have more Significant effect.

Where is our data from?

We get the data from the Ministry of Transportation and Communications R.O.C, the data is official data which is credibility.

How to do?

1. Buildng the model



After cleaning and dummying the data, since the number of high risk is lesser compare with low risk, it will make the high risk be determined as outlier, we do the oversampling to overcome this problem.

And we use a few method, compare them with cost matrix to find the best one.

2. we have to collaborate with government to get the information of the rider. (gender, violation, the purpose of using bicycle, job).
3. put the data of step2 into the model.

Recommendation

we suggest future practitioners to do:

- Improve the misclassification of whom is not victim but be predicted as.
- Connected these data to the actual death in the future. Track the responder of survey and combined them to the results in the future (slightly/badly car accident/ death in the accident).

1. Problem

Business goal

Traffic accident happens almost every day, and it is more risky for a motor-cycle rider for a car driver. Besides the death or mutilation, the social cost (ex: medical expenses, traffic jam) is the main cost caused by the car accident. In order to lessen the cost, we design this model. And we view the government as our customer because it has the most powerful resource and pursue the social benefit maximization.

Dataming goal

It is a supervised predictive task. We develop classifications for high / low risk groups. And because we want to predict who will be the high risk group, so our goal is predict.

2. Data description

After exploring and cleaning the dataset, we only keep 7208 records including 11 predicators and 1 outcome variable in our model. The data sample is displayed below.

A record represents a responder of daily motorcycle usage survey, provided by ministry of traffic.

Input variable:

Capacity: Ordered categorical variable
Number of passengers per ride on average.
Purpose: Categorical variable
Main purpose of using the motorcycle.
Use day per week: Ordered categorical variable
Number of rides in a week on average
Average duration: Ordered categorical variable
Duration per ride on average
Violation: Ordered categorical variable
Number of violation times in the last year (self-reported)
Gender: Categorical variable
Male/Female (4228/2980)
Age: Categorical variable
Age of the responder
Education: Ordered categorical
Education level of the responder
Job: Categorical

Job type of the responder

Income: Ordered categorical

Income interval of the responder

City of motorcycle: Categorical variable

Where is the motorcycle of the responder located (We will end up turn it into area of motorcycle variable)

Output variable: accident (high/low risky).

Variable we want to predict. We assume that high risky accident wrecker will engage in more dangerous injury and cost more social cost. They are objectives whom government should take some actions to.

capacity	purpose	use day pe	average du	violation	gender	age	education	job	income	city of motorcycle	area of mo	accident(low&high)
1	5	1	2	0	0	3	6	13	6	1	1	0
1	4	7	3	0	0	5	1	22	1	3	2	0
1	5	3	2	0	1	5	4	13	6	8	1	0
1	1	1	3	0	1	6	4	22	3	8	1	0
1	1	7	3	0	0	3	3	12	3	8	1	0
1	5	7	2	0	0	6	5	15	5	22	5	0

Figure sample of original data after cleaning

3. Data Preparation

In data preparation, we start with cleaning records that do not have our output variable (accident). We deleted 3273 records (10481-7208) and have 7208 records for our analysis. After exploring the data, we found the responded average accident per year is skewed. The data are mainly concentrated on option zero, one or two times. On the other side, the answered options of more than two times per year are relatively scarce (695/ 7000). Therefore, we separate them (who answered more than two) as high risky group, denoted by 1, who are the targets we want to capture, and those less than two are grouped as low risky, denoted by 0. Under this condition, using lift chart will be more relatively complicated. So we choose to use the classification, trying to find the high class and low class, which will have enough data to know the characteristic of each group. We also transfer one of our input variable, city of motorcycle, into area of motorcycle. We successfully reduced 22 dimensions into 5 dimensions. Next, we create dummy variables for our nominal variables (purpose, job and area of motorcycle). Finally, we accomplished all the work in our data preparation and have 7208 rows and 31 columns of data to do the analysis.

4. Data mining solution

First of all, we partition our data into training, validation and test three parts. In our data, we

have very few records of high risky group (1.86%). Therefore, we did a data partition with over-sampling and let the training data set has fifty percent of high risky group records. We used XLminer to run a classification tree to select important variables. We choose first three layers of the full tree, and found out that violation, gender, usage per week, purpose, job and income is the most important variables. Next step, we use these variables to run Naïve Bayes, K-Nearest Neighbors, Neural Network and Logistic Regression and set the cutoff probability to 0.5. We use confusion matrix as our performance evaluation. From the result table, we can see that Logistic Regression capture high risky group more accurate than other methods and KNN provide the smallest overall percent of error. To interpret our result more specifically, we create a cost matrix to compute the final cost of each method. According to government report, death rate in traffic accident is 15%and per death increase cost 15,720,000 NTD, while per injured increase 1,190,000 NTD cost. We assume that the cost of our strategy on high risky group is 1,000 NTD per case. After we consider the cost matrix me made, Logistic regression giving the best result, which means that it can save more money than other models. So we choose Logistic Regression model as our final data mining solution.

4. Conclusions

Some Rising Issues of Business goal

By using our model, we provide a convenient way to identify high risky group for government to take some actions to prevent them from dangerous accident and expend heavy social cost. However, there're some serious emerging issues. For example, who have authority to browse and use such output, and what would people feel and think about when they are defined as high risky group.

We think there's still a call for figuring out better utilization of the data output.

Limitation and Recommendation

In our model and evaluation, we assume the cost of transforming a high risky rider to low risky rider is much less than the social cost expenditure. However, in reality, things would not go that perfect. For example, we may have larger cost to change riding behavior of one high risky rider. As a result, our performance evaluation may have to change the cutoff value to find the most optimized models that does not misclassify that much, and not merely try to capture more high risky riders as we've done in this study. Based on the conclusions above, we suggest future practitioners to do:

- Improve the misclassification of whom is not victim but be predicted as.
- Connected these data to the actual death in the future. Track the responder of survey and combined them to the results in the future (slightly/badly car accident/ death in the accident).

Appendix

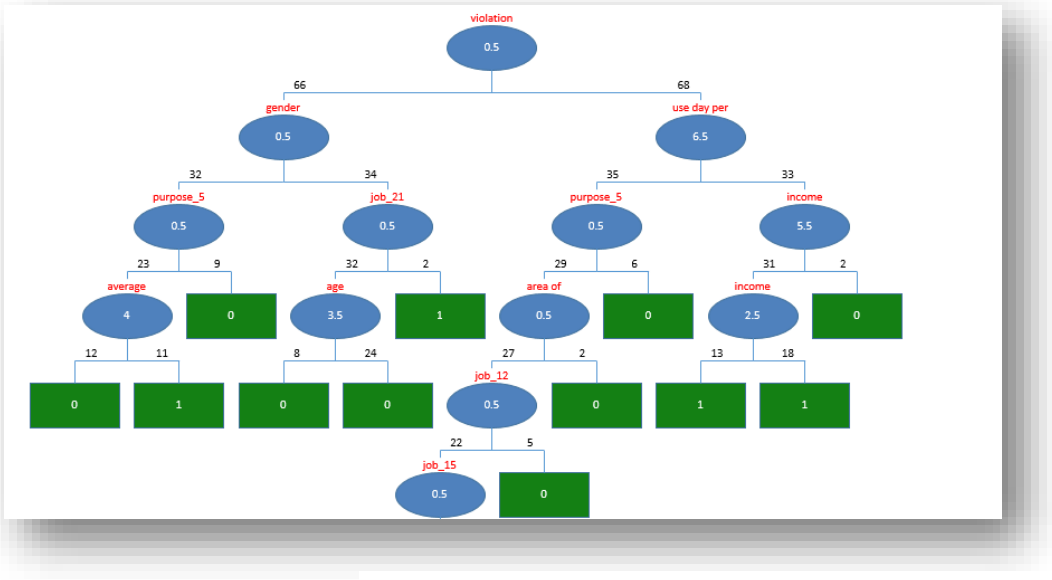


Fig1. Classification Tree

Classification Confusion Matri:

	Predicted Class	
Actual Class	0	1
0	1174	620
1	9	25

Error Report

Class	# Cases	# Errors	% Error
0	1794	620	34.5596
1	34	9	26.4706
Overall	1828	629	34.4092

Naïve Bayes

Classification Confusion Matrix

	Predicted Class	
Actual Class	0	1
0	1454	340
1	22	12

Error Report

Class	# Cases	# Errors	% Error
0	1794	340	18.9521
1	34	22	64.7059
Overall	1828	362	19.8031

KNN

Classification Confusion Matrix

	Predicted Class	
Actual Class	0	1
0	1128	666
1	8	26

Error Report

Class	# Cases	# Errors	% Error
0	1794	666	37.1237
1	34	8	23.5294
Overall	1828	674	36.8709

Neural Network

Classification Confusion Matri:

	Predicted Class	
Actual Class	0	1
0	1216	578
1	7	27

Error Report

Class	# Cases	# Errors	% Error
0	1794	578	32.2185
1	34	7	20.5882
Overall	1828	585	32.0022

Logistic Regression

Fig2. Result Table