Improving Road Safety
by Profiling Different Accident Type

Business Analytics Using Data Mining

Team7

Angela Hung

Aylada Khunvaranont

Celia Chen

Dobby Yang

Mahsa Ashouri

Executive Summary

Business Problem We all are road users and everyday we see the news about car accidents, which leads us to the concern of what can Transportation Department of Taipei City Government do to decrease the number of the accidents. Therefore, our project focuses on how to identify what factors or conditions have influence on the certain accident types. With our findings, the government can pay more attentions to those area where accident happened. Moreover, the government can manage their budgets wisely and effectively at where they will fix what; for instance, fixing the road conditions, install more equipment, install more traffic lights and etc.

Data The raw data we get is from Data Taipei (http://data.taipei/) which has data that is collected by government sectors and is a reliable source. We have combined three years of data set into the data that we will use to analyze.

Analytics Solution There are several combinations of different software, algorithm and data preparation methods taken to reach the best solution of this business problem. After we tried, decision tree grown by RapidMiner is chosen as the final solution. Although it is not the best-performance, it has the most powerful explanatory ability, which is the most important element in explaining data field.

Recommendations We found seven exact conditions of *BackHit*, *SideHit*, *Ped_crossing* and *Scratch* by profiling the data we had. According to the result, we are able to provide some advice for the Transportation Department of Taipei City Government to improve the road safety efficiently.

Detailed Report

Problem Description

Business Goal

Since we have seen news about traffic accidents everyday and it is what we need to take caution for in our daily lives, we would like to assist Transportation Department of Taipei City Government to improve the road safety in Taipei with more efficient way by learning the conditions that distinguished different type of accidents.

With our results, we can help the Government to initiate fixing road projects as in fixing the road condition or install more equipments, such as traffic lights, caution signs and so on, that can help reduce the number of particular accident types. Moreover we aim at helping government for managing their budget wisely and efficiently. We can suggest them where and when can be best time for spending their budget. Taipei citizens would benefit from this in the sense that the road using becomes safer.

Data Mining Goal

To achieve the stated business goal, we will analyze previous accident data and identify the factors that differentiate each accident types. We are going to profile these accident types by identifying the factors that might be the cause of that accident. This is a supervised, descriptive, and ongoing project.

Data Description

We have obtained data set of accident records in Taipei from year 2011-2013 from Data Taipei (http://data.taipei/). First of all, we have combined the records that have the same case ID as one accident case with one record. As a result, we have total of 54,797 rows and 11 columns from three years data set (2013: 17,991rows; 2012: 17,843 rows; 2011: 18,963 rows).

Our input variables are categorized as time variables and environment variables. For time variables, we have included *Hour* (by Daytime or Nighttime), *RushHour*, and *Weekday*. In environment variables, we have *Weather*, *Light*, *Speed_Limit*, *Road_type*, *Acc_location*, *Pavement_condition* and *Signal*.

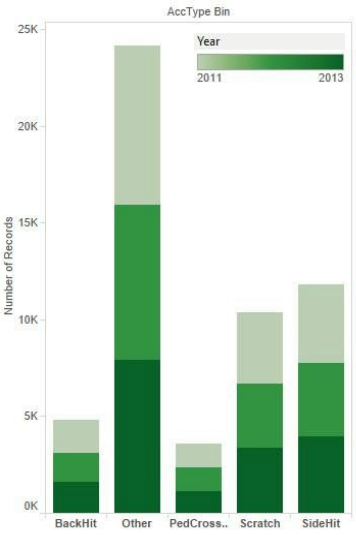
For output variables, we have binned into five categories according to what we think it happens a lot or causes serious damages, which are *SideHit*, *BackHit*, *Ped_Crossing*, *Scratch*, and *Others*.

CASE_NO	YEAR	Hour_Bin	RushH	Weekda	Weatl	Light_	Speed_	Road_	Acc_location	Paven	Signal	AccType_Bin
C5103031	2013	DayTime	NoRush	WeekEnd	Sunny	Light	50	Lane	MotorcycleLan	Wet	None	BackHit
C5103033	2013	DayTime	NoRush	WeekEnd	Sunny	Light	50	Lane	NormalLane	Wet	None	BackHit
C5103034	2013	DayTime	NoRush	WeekEnd	Sunny	Light	50	Lane	FastLane	Wet	None	Other
C5103035	2013	DayTime	RushHo	WeekDay	Sunny	Light	50	Lane	MotorcycleLan	Wet	None	Scratch
C5103037	2013	DayTime	NoRush	WeekDay	Sunny	Light	50	Fork	CrossRoad	Wet	TrafficP	Other

Table 1: Example of our cleaned data

Data Preparation

Originally, we had 119,721 rows and 30 columns (2013:39,577 rows, 2012:39,062 rows, 2011:41,082 rows). In data preparation, we have deleted records that are the accidents happened on highway, which is not our goal. We also binned the categories that are similar and deleted some columns that are not significant in profiling accident types in order to reduce the dimensions. Moreover, we derived new columns of Hour by Daytime or Nighttime and Rush hour or not from the original Hour column.



After cleaning our data, we try some visualization. In Figure 1, we can see that our output variables are imbalanced. The category "*Other*" gained the biggest proportion in our data set. So to deal with this problem, we tried the oversampling and undersampling methods to our data in order to have the balance and unbiased data, which we called it oversampled data in the following.

We have applied oversample/undersample to our data set by setting each output category to have equal records of 5,000; in total, we have 25,000 rows of data to run our algorithm.

Figure 1: The proportion of each accident type in different year

Data Mining Solution

To deal with this profiling task, some of the data mining methods that we've considered include decision tree, logistic regression and discriminant analysis. We have settled down with the first two methods in this research due to their strength of explaining the results. The analytical tools we used in this project are RapidMiner and R.

We have tried both original and oversampling data to train our models to see if the model trained by oversampling data works better. Logistic regression performs better without oversampling, whereas decision tree performs better with oversampling in terms of overall accuracy[1]. We then adopted the result of decision tree from Rapidminer as our final solution because of its high readability and interpretability.

The RapidMiner process flow and the tree we have generated is shown in appendix. Table 2 shows the important factors that we have identified with our algorithms by their ranking of importance. Note that both algorithms agree that *Acc_location* and *Speed_limit* are the key factors that distinguish between accident types, other factors include *Road_type*, *RushHour*, *Weather* and *Weekday*.

[1] Here's an issue with oversampling. Whether oversampling should be applied in a profiling task is debateful.

Decision Tree		Logistic Regression	
1	Acc_location	1	Acc_location
2	Speed_limit	2	Speed_limit
3	Road_type	3	RushHour
4	Weather	4	Weekday

Table 2: Variables that distinguish different accident types

Outcome	Factors
BackHit	1. Speed limit > 75 2. at motorcycle lane 3. Speed limit > 65 , not at cross road
SideHit	1. Speed limit > 65 at cross road
Pedcrossing	1. at zebra crossing
Scratch	1. Speed limit < 65, other road type, not cloudy, motorcycle lane. 2. Speed limit > 65, not at cross road, rainy.

Table 3: Decision rules of accident types of our interests

Table 3 shows the decision rules we concluded from the decision tree. They are ranked by their significance in terms of the number of records in the end nodes (leaves).

Conclusion

Through our results, we found that under most of the circumstances, all kinds of the accidents could happen. We are only able to successfully distinguish about 5% of the accidents, and generate some decision rules for the accident type of our interests according to this 5% of the data. However, if we are able to identify some circumstances that could cause certain types of accidents, this could be helpful. We have also learned some important factors that distinguish between accident types. With further assistance from road traffic expertise, treatments and improvements could be applied to reduce certain types of accidents.

There are some issues that are raised in this project and should be more carefully studied in the future such as: What are the proper ways to measure the performance of a profiling task? Is the classification accuracy the only thing to be considered? Is the oversampling technique suitable in a profiling task? This study could be improved if these issues are handled more carefully.

Appendix

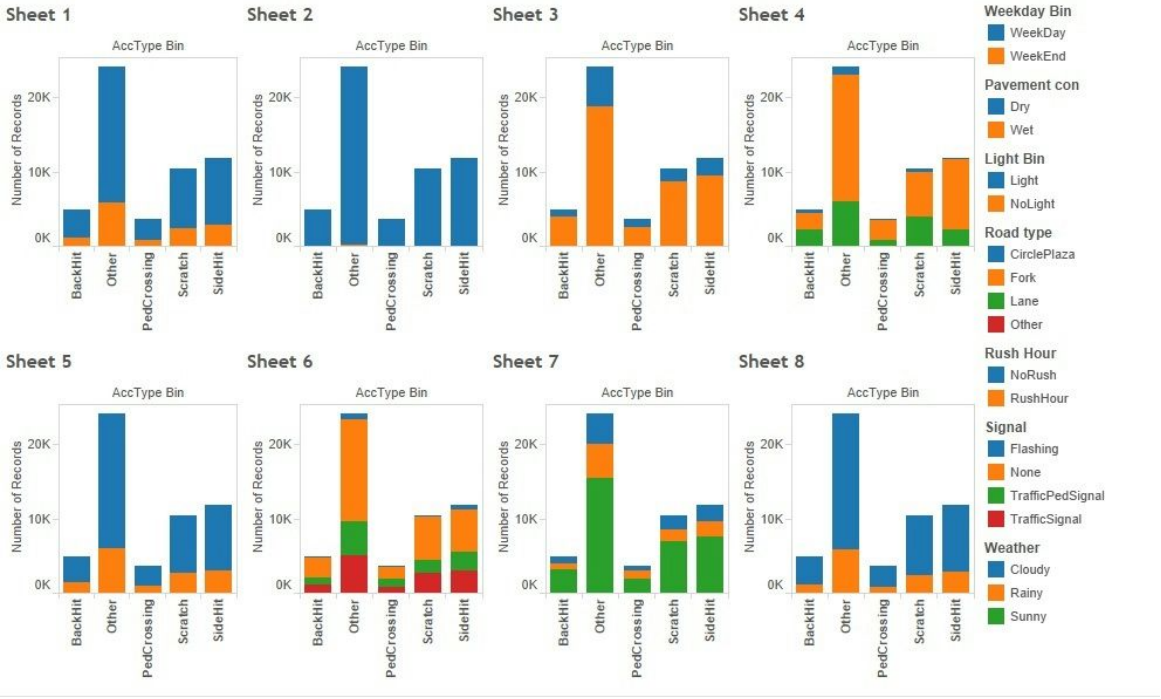


Figure 2: The proportion of each accident type by using each of our input variables

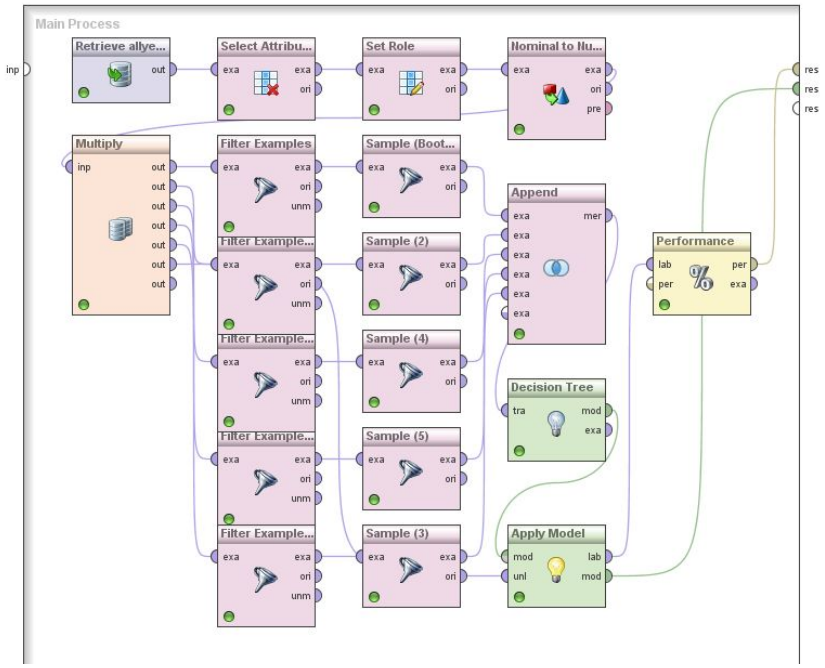


Figure 3: RapidMiner Process Flow for our model

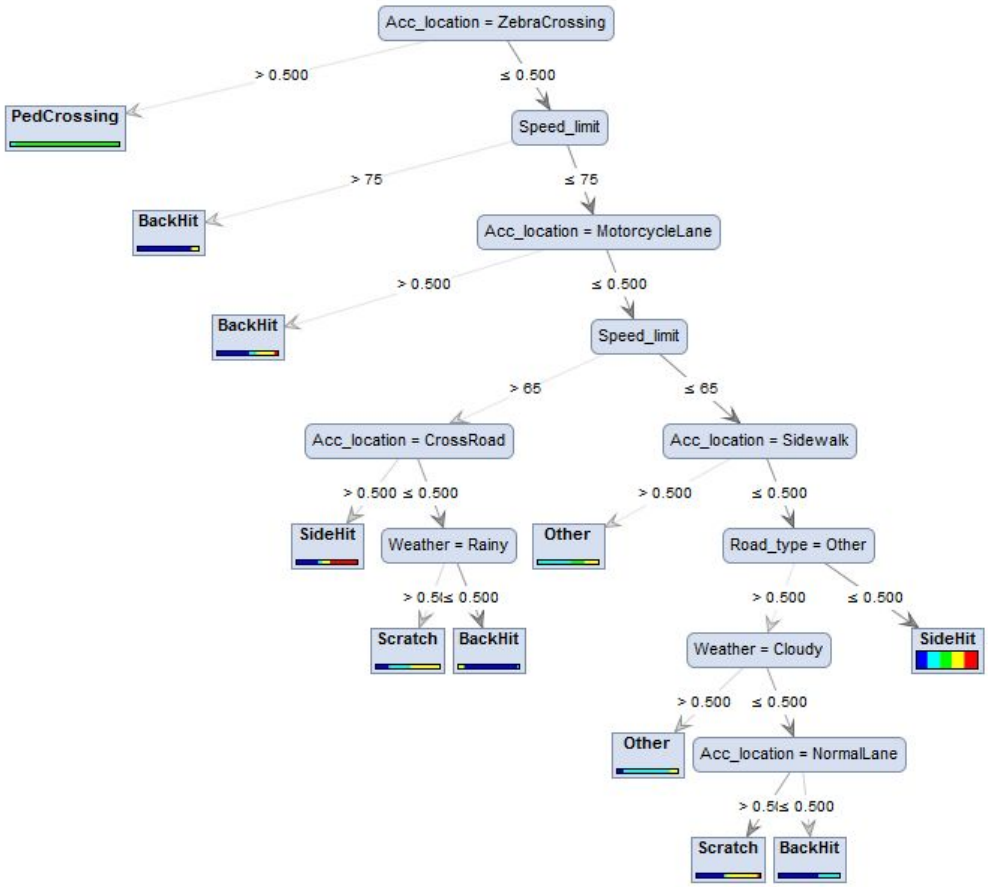


Figure 4: Decision Tree generated from RapidMiner

True/Pred.	SideHit	Scratch	Other	BackHit	PedCrossing	class recall
SideHit	11749	2	5	65	0	99.39%
Scratch	10051	19	3	312	0	0.18%
Other	23670	24	60	334	82	0.25%
BackHit	4432	17	2	371	0	7.69%
PedCrossing	3327	0	5	8	259	7.20%
class precision	22.07%	30.65%	80.00%	34.04%	75.95%	

Table 4: Confusion matrix of the decision tree

3 years without FrontHit, orinigal data								
Rref \ Pred	BackHit	Other	PedCross	Scratch	SideHit		overall Acc	overall Acc-other
BackHit	226	4234	0	362	0	0.046869	54797	30627
Other	164	23512	81	413	0	0.972776	24413	901
PedCrossin	8	3282	257	52	0	0.071409	0.44551709	0.029418487
Scratch	125	9842	0	417	1	0.040154		
SideHit	44	11544	0	232	1	0.000085		
	0.39859	0.44858	0.76036	0.28252	0.5			
3 years without FrontHit, overundersample data								
Rref \ Pred	BackHit	Other	PedCross	Scratch	SideHit		overall Acc	overall Acc-other
BackHit	2385	316	797	238	1086	0.494608	54797	30627
Other	5069	4638	5275	968	8220	0.191891	14017	9379
PedCrossin	609	347	1425	117	1101	0.395943	0.25579868	0.306233062
Scratch	3874	799	1835	525	3352	0.050554		
SideHit	2063	1519	2756	439	5044	0.426698		
	0.17036	0.60874	0.11789	0.22956	0.26826			

Table 5: Confusion matrix of Logistic Regression