Predicting the Risk Level of the Vehicle for Customer's Auto Insurance Using JMP® Pro 10

Dinesh Yadav Gaddam, Musthan Kader Ibrahim
Management Information Systems, Oklahoma State University, Stillwater, OK USA

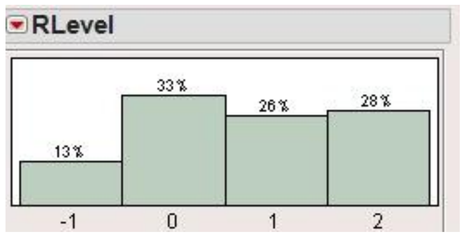
Introduction

Relating used car risk to vehicle characteristics is often a challenging task but it is very important for figuring out used car insurance premium. This poster is an attempt to build a model that uses vehicle characteristics as inputs and an overall measure of risk for a used vehicle as a target with the help of SAS JMP Pro 10.
This would help insurance organizations in deciding the more profitable customers and to quote proper insurance premiums for the risky vehicles, therefore increasing their net revenue through them.

Data Preparation

The data used for this purpose contains two types of entities, one having the specifications of an auto in terms of its various characteristics like make, fuel-type, curb-weight, horse power, peak-rpm etc., and the another one contains its assigned insurance risk rating(Rlevel) which is being mapped according to its price and later it is being adjusted by moving up or down along the scale. The data has about 264 observations and collectively gathered from Insurance Collision report – Insurance Institute for highway safety.

The Risk level of the auto is being spread across five levels from -2 to +2. The value -2 indicates the auto is too risky and +2 being very safe. The number of cases with the risk level -2 are very less so these cases are merged into next risk level -1. Therefore we have the four risky levels -1,0,1 and 2. Most of the categorical variables which have more number of levels have been consolidated into few levels. For an instance a total of 21 Car manufacturers have been identified in the collected data and they have been consolidated to four groups Grp1 to Grp4 which helped in predicting a better model.



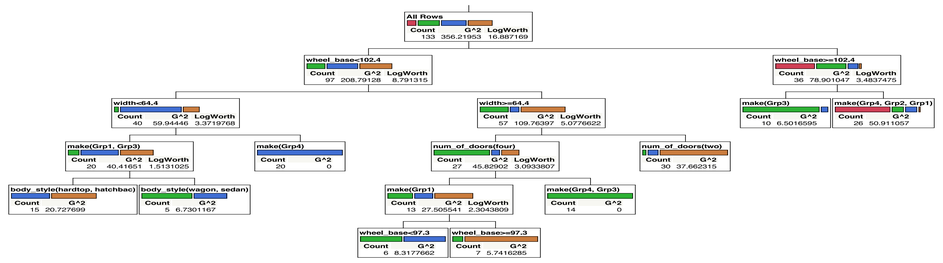
Modeling

The target variable in this analysis is Nominal. So it is suggested to use Misclassification rate to find out the best model. The data has been split into 70% training data and 30% Validation data. The data set contains 24 input variables and one target variable. The important variables have been identified which help in predicting a better model. Decision Tree, Neural Network, Logistic regression and Bootstrap forest models have been built using extensive visual and graphical capabilities of JMP Pro 10.

Among all the models being built, Decision Tree model has the least **Misclassification Rate** indicating that it has the high correctness while predicting the occurrence of an event. It calculates the proportion of an observation being allocated to the incorrect group. The below table represents the Misclassification Rates with respect to the models being considered. Misclassification Rate = Number of incorrect Classifications / Total Number of classifications.

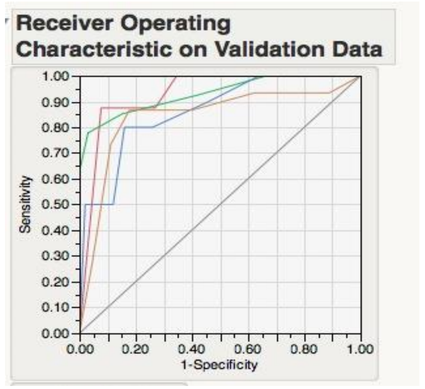
Model	Validation Data	
	Misclassification Rate	RMSE
Decision Tree	**23.33%**	42.15%
Neural Network	33.33%	54.51%
Bootstrap Forest	46.30%	64.26%
Logistic regression	29.02%	45.66%

Graphical representation of the Decision Tree



Results

The below visual graphs represent area under the ROC curve is the proportion of these pairs (Sensitivity vs. 1-Specificity) in which the posterior probability is higher for the event case. Thus the Area Under the Curve values(~0.9) shown below for different levels from decision tree implies that all event cases have higher posterior probabilities than all non-event cases in their specific risk levels.



One away accuracy has been calculated and shows the Decision Tree model has done a good job on the Validation data. One away accuracy is the ability of the model to make a prediction which does not deviate from the actual value by more than one class.

Risk-Levels	-1	0	1	2
One Away Accuracy	82%	85.12%	91.12%	93.33%

Segmentation analysis has been done and the following table displays the Important auto characteristics with respect to the each Risk level of an auto.

Segment	Most important predictive variable
2	Height, Wheel_Base, Curb_Weight
1	Width, Wheel_Base, Curb_Weight
0	Horsepower, Wheel_Base, Height
-1	Wheel_Base, Length, Make

Conclusions

• Wheel Base Length is an important key driver when deciding an Auto to be risky or not. Wheel base length with more than 102.4 inches seems to be more risky than otherwise.

•Wheel base length less than 102.4 inches and width > 64.4 inches are more safer auto vehicles than the vehicles with width < 64.4.

•Vehicles which fall under Grp-3 (BMW, Honda,Subaru, Peugot) are less risky vehicles than the rest.

Reference

- http://www.jmp.com/support/help/Modeling_and_Multivariate_Methods.shtml#1034510
- http://www.jmp.com/applications/modeling/

Acknowledgement

We would like to thank Dr. Goutam Chakraborty Professor (Marketing) and founder of SAS and OSU Data Mining Certificate Program - Oklahoma State University for his generous support throughout this project.