|  |
| --- |
| Comparison of Bayesian Networks |
| Abhinaya, Allen, Aravind, Kavya, Preethi Jennifer R |

# Introduction

We have taken the Vehicle Safety dataset and tried to predict the severity of injuries in car accidents. For this purpose, we’ve considered the models; Naïve Bayesian and Tree Augmented Network.

**Business Objective:**

In addition to predicting the severity of the injury, as a business objective, we would also like to compare the models to confirm which of the two can predict the fatality of an injury in a vehicle accident better. From the data, we have considered OA\_MAIS = 5 & 6 to be fatal.

**Naïve Bayesian:**

Naïve Bayesian is the simplest classification technique which assumes mutual independency of the features.  It is useful for large datasets and is based on the Bayes’ theorem:

P(A|B) = P(B|A) \* P(A)/P(B)

**Tree Augmented Naïve Bayesian:**

In real time applications, it is unlikely that complete independency between the features exists. TAN accounts for partial dependency of the features where it assumes that each variable depends on the class and one other variable.  TAN can be regarded as a semi naïve Bayesian network.

For our assignment, we will be comparing the accuracy of each of the model and identify the optimal model of the two for the given dataset.

# Method

A dataset consisting of 20247 observations, 20 independent variables and a dependent variable were taken for the Bayesian Network model.

DEPENDENT VARIABLE:

* Maximum Known Occupant AIS i.e. Injury Severity (OA\_MAIS)

INDEPENDENT VARIABLES:

* Vehicle Curb Weight (GV\_CURBWGT)
* Lateral Component of Delta V (GV\_DVLAT)
* Longitudinal Component of Delta V (GV\_DVLONG)
* Energy Absorption (GV\_ENERGY)
* Number of Lanes (GV\_LANES)
* Vehicle Model Year (GV\_MODELYR)
* Weight of the Other Vehicle (GV\_OTVEHWGT)
* Speed Limit (GV\_SPLIMIT)
* Truck Weight Code (GV\_WGTCDTR)
* Age of Occupant (OA\_AGE)
* Air Bag System Deployed (OA\_BAGDEPLY)
* Height of Occupant (OA\_HEIGHT)
* Manual Belt System Use (OA\_MANUSE)
* Occupant’s Sex (OA\_SEX)
* Occupant’s weight (OA\_WEIGHT)
* Deformation Location (VE\_GAD1)
* Average Track Width (VE\_ORIGAVTW)
* Wheelbase (VE\_WHEELBAS)
* Clock Direction for Principal Direction of Force (VE\_PDOF\_TR)
* Vehicle Footprint (GV\_FOOTPRINT)

## Data Cleaning

Almost 47 % of the data had missing values split across several variables. We adopted a 4-step criterion for data cleaning

**Step 1: Removing variables**

It was decided that variables with more than 40 % of missing values were to be removed from the dataset. However, no variable fell under this criterion.

**Step 2: Removing missing value from dependent variable.**

The missing values in the dependent variable account for only 5% of the entire data. For validating our model prediction against the actual dependent variable, all the observations with the missing values for the dependent variable were removed.

**Step 3: Handling missing values of variables with continuous data**

In order to have a least effect on the final result, the average of each variable was used to replace the missing values with the exception being VE\_PDOF\_TR .

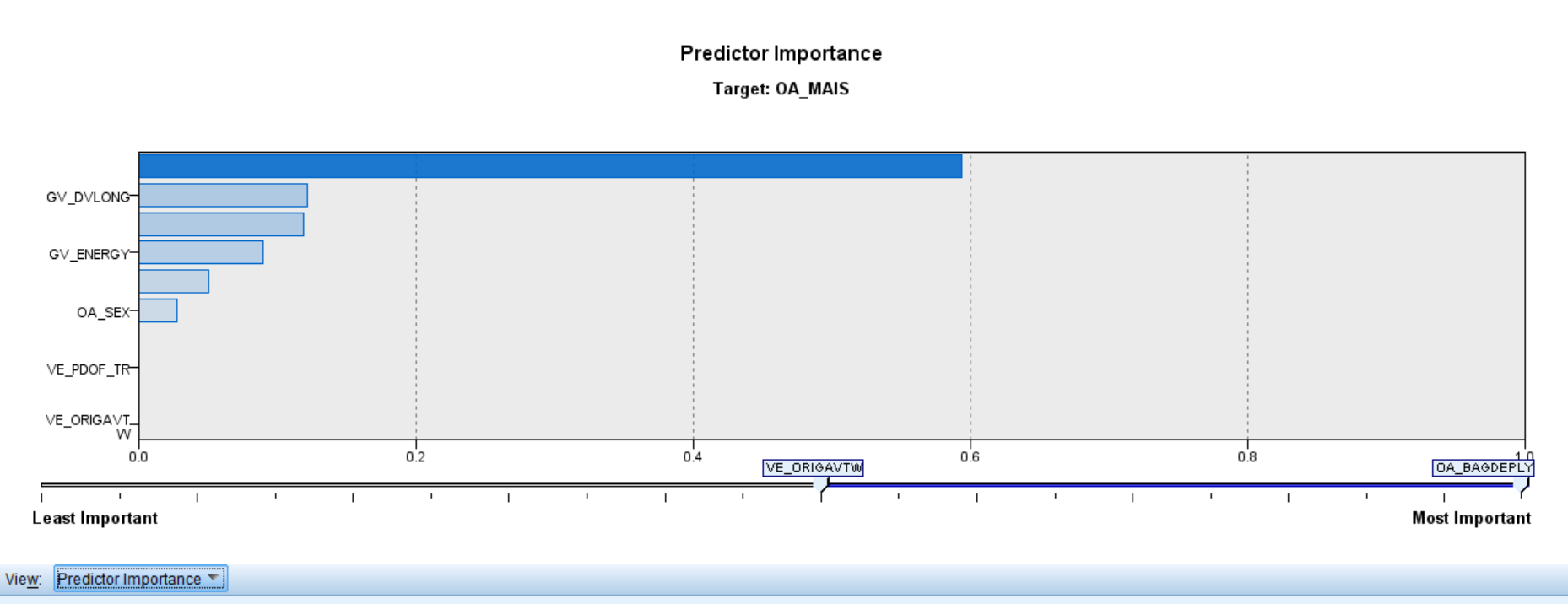
VE\_PDOF\_TR refers to the direction of the highest force acting on the vehicle with the collision directions ranging between 0 to 355 in degrees of 5. Hence to resonate with the remaining data the average (152) was rounded up to 155 and this value was used to replace missing value.

**Step 4: Determining how to handle missing values of variables with discrete data**

However, when faced with discretized data, averaging was not possible hence we had to choose between removing the observations of these variables with missing data or replacing them with the mode.

The two variables in question are OA\_SEX and VE\_GAD1.The missing observations of these 2 variables together accounted for 5 % of data. Removing these observations may result in significant loss of data. It was decided that if the relative importance of the two variables in determining the response variable was high, their missing value observations would be removed to not skew the prediction with inflated representation of their mode. However, if their relative importance is low, we could afford to replace the missing values with the mode with minimal distortion of data while retaining significant information.

Using SPSS modeler, the predicted importance of each variable was measured to identify the relative importance of the two variables in question.



As per the graph above, both variables have the least importance. Hence to avoid the loss of valuable data from other variables, the modes of each variable were taken to replace the missing value as opposed to deleting the observations.

## Discretization

We have adopted a simple criterion for binning wherein we choose the bin size (between 5 to 10) which has the smallest ratio of the highest count to lowest count in the bin.

|  |  |
| --- | --- |
| VARIABLE | BIN SIZE |
| GV\_CURBWGT | 8 |
| GV\_DLAT | 10 |
| GV\_DVLONG | 5 |
| GV\_ENERGY | 5 |
| GV\_LANES | 7 |
| GV\_MODELYR | 10 |
| GV\_OTVEHWGT | 10 |
| GV\_SPLIMIT | 7 |
| OA\_AGE | 7 |
| OA\_HEIGHT | 8 |
| OA\_MAIS | 7 |
| OA\_MANUSE | 2 |
| OA\_WEIGHT | 9 |
| VE\_ORIAVTW | 8 |
| VE\_WHEELBAS | 10 |
| VE\_PDOF | 8 |
| GV\_FOOTPRINT | 8 |

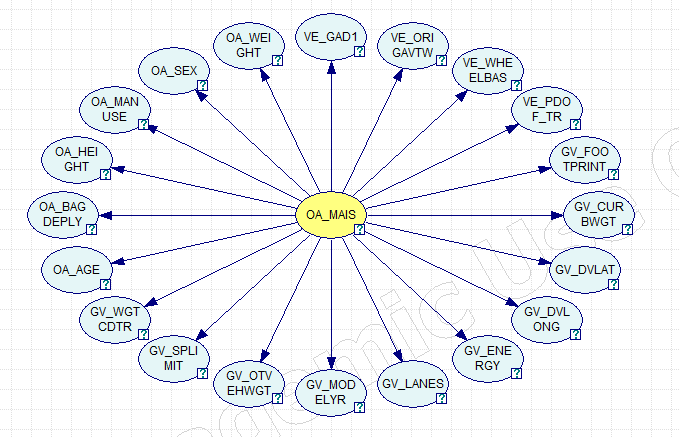
NOTE: Nominal alphanumeric variables (GV\_WGTCDTR, OA\_BAGDEPLY, OA\_SEX, VE\_GAD1) were not binned and OA\_MANUSE was binned into 2 since the column was assumed to be binary.

## Splitting Data

The discretized data was split in the ratio of 75:25 for training and test data respectively using JMP tool.

# Naïve Bayes Network

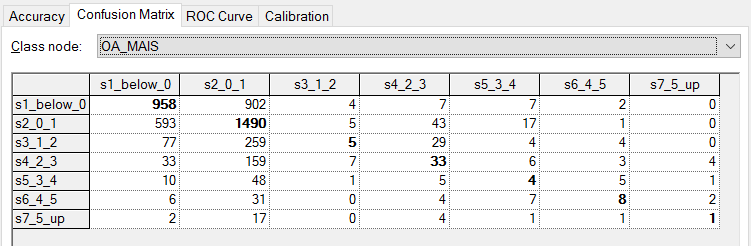
Naïve Bayesian network was created using the training data setting OA\_MAIS as the class.



As can be seen from the above network, the dependent variable is related to all the independent variables however none of the independent variables are related.

# Validation of Naïve Bayesian Network

The test data was validated for the modeled network resulting in the following output:



The model had an overall accuracy of 51.95 %.

For serving the business objective mentioned earlier, the confusion matrix of s7\_5\_up was taken as a severity of 5 and 6 are considered to be fatal. Its accuracy, precision and recall were calculated.

Accuracy = (True Positive + True Negative) / Total number of predictions

Precision = True Positive / Predicted Positive

Recall = True Positive / Actual Positive

|  |  |  |  |
| --- | --- | --- | --- |
| Naïve |  | Predicted | |
|  |  | S7 | Not S7 |
| Actual | S7 | 3 | 16 |
| Not S7 | 61 | 4633 |

Here,

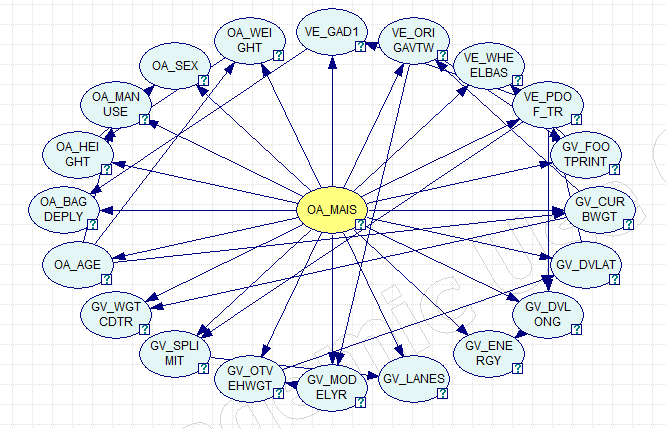
Accuracy = 98.4 %

Precision = 4.7 %

Recall = 15.79 %

# Tree Augmented Network

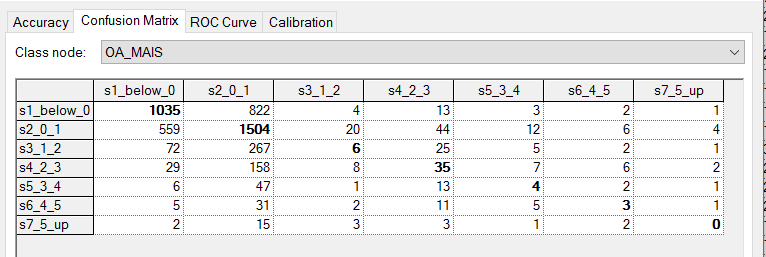
Tree Augmented network was created using the training data setting OA\_MAIS as the class.



It can be seen from the graph that the dependent variable is related to all the independent variables and in addition there is a partial dependency between the independent variables.

# Validation of Tree Augmented Network

The test data was validated for the modeled network resulting in the following output:



The model had an overall accuracy of 53.78 %

For serving the business objective mentioned earlier, the confusion matrix of s7\_5\_up was taken and its accuracy, precision and recall were calculated.

Accuracy = (True Positive + True Negative) / Total number of predictions

Precision = True Positive / Predicted Positive

Recall = True Positive / Actual Positive

|  |  |  |  |
| --- | --- | --- | --- |
| TAN |  | Predicted | |
|  |  | S7 | Not S7 |
| Actual | S7 | 0 | 19 |
| Not S7 | 7 | 4687 |

Here,

Accuracy = 99.4 %

Precision = 0.0 %

Recall = 0.0 %

# Inference:

From the validation, we have taken two considerations.

1. Compare the two models’ overall accuracy:

It can be seen that the Tree Augmented Naïve Bayes (TAN) has a better prediction accuracy when compared to Naïve Bayes. This is in tandem with the nature of TAN which accounts for partial dependencies of the features. There is a marginal dependency among the features, hence TAN would be a better network modeling technique of the two.

Thus, while comparing the two models, one would choose the Tree Augmented Naïve Bayes (TAN) model.

1. Evaluate the models for the purpose of our business objective:

To select a model that better suits our business objective, more importance has been given to the recall percentage as it stands for the percentage of the total actual positives that have been predicted correctly and we would want a model that can better capture all the fatal records.

When comparing the two models it can be seen that while the recall percentage is very low in both, it is worse in TAN with a recall of 0%. A reason for this could be because of the limited number of observations available in this bin (in the training set) making this a rare event which is hard to predict. In order to improve the recall percentage, we suggest that more data that falls under this bin (OA\_MAIS = 5 & 6) be collected.

# Conclusion:

1. Compare models:

For any model, creating lesser no of bins will inherently result in more prediction accuracy, however the prediction itself will be for too broad a range (of individual bins) to provide significant information. Thus, there exists a tradeoff between bin sizes and prediction accuracy.

1. Evaluate the models for the purpose of our business objective:

In the case of prediction for a rare event, the quantity of the data collected should be more for any model to predict with a high recall.