**PREDICTING INJURY SEVERITY OF DRIVERS**

MSIS 5633 – Business Intelligence Tools and Techniques

Team: Team 10

Members: Darshini Priya (A20108860)

Suhao Chen (A20164586)

Arjun Achuthan (A20115366)

December 2018

Contents

[Introduction 2](#_Toc532686481)

[CRISP – DM methodology 2](#_Toc532686482)

[Business Understanding: 3](#_Toc532686483)

[**Business Objective:** 4](#_Toc532686484)

[**Data Mining Goal:** 4](#_Toc532686485)

[**Project Planning:** 4](#_Toc532686486)

[Data Understanding 5](#_Toc532686487)

[DATA PREPROCESSING: 5](#_Toc532686488)

[Descriptive Statistics: 10](#_Toc532686489)

[MODEL BUILDING: 12](#_Toc532686490)

[Model Evaluation/validation 18](#_Toc532686491)

[Deployment/ Suggestions: 19](#_Toc532686492)

## Introduction

Every year we have a huge number of road accidents happening in the US. Thus, it needs a thorough understanding of the circumstances and factors that lead to an accident and the resulting injuries to be able to improve the situation. For this we use the car crash data provided by NHTSA and leverage our data mining skills to analyze the data and suggest a suitable data mining model that can best predict the injury severity of the drivers.

Further, based on the analysis we will suggest recommendations to avoid severe injuries during accidents.

# CRISP – DM methodology

CRISP-DM stands for cross-industry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project. It's an open standard for anyone to use, which was developed by cooperation with over 200 organizations. It was developed specifically for data mining; however, it is flexible enough to be applied to many analytic styles.



The CRISP-DM process model has six major phases:

* Business Understanding: To have a good understanding of the business problem that is to be solved and addressing the goals and objectives.
* Data Understanding: Inspect the data and gain a better understanding about the data to evaluate them.
* Data Preparation: Modify data to a state such that it is fit for analysis.
* Modeling:  Develop analytical techniques to develop a good model that can be presented to the business.
* Evaluation: Compare the models and find the best one that can be implemented.
* Deployment:  Present the final model to the Business such that it can be integrated to everyday business.

## Business Understanding:

We focus our research on the driver’s injury severity. Car crash injury severity of drivers is probably related with the following potential factors:

* Drivers’ control ability, e.g. whether drivers drink alcohol or take drugs;
* Safety restriction usages, e.g. the proper use of safety belts, airbags, crash avoiding systems, etc.
* External conditions, such as the weather, road conditions, etc.
* Crash types and how the drivers react to the accidents.
* Driver demographics, such as sex, age, etc.
* Other factors.

Our analysis is based on the GES database from NHTSA. About NHTSA and their Mission:

* NHTSA collets car crash data to support its mission to reduce motor vehicle crashes, injuries and deaths on our National Highways and roads.
* They use data from many sources which includes National Automotive Sampling System (NASS) and General Estimates System (GES).
* The National Automotive Sampling System (NASS) - General Estimates System (GES) data are obtained from a nationally representative probability sample selected from all police-reported crashes.

### **Business Objective:**

* To identify the key factors that contribute towards the injury severity of drivers.
* To provide suggestions and recommendations based on the insights from our analysis.

### **Data Mining Goal:**

* To identify variables that contribute most to the injury severity of the Driver using analytical methods.
* Building a suitable analytics model with the available data to predict likelihood of injury severity.

### **Project Planning:**

|  |  |  |  |
| --- | --- | --- | --- |
| Phase | Time | Resources | Risks |
| Business understanding | 1 week | Analysts, Analytical user manual | Time, Lack of understanding |
| Data understanding | 1 weeks | Analysts | Data problems, technology problems |
| Data preparation | 2 weeks | Analysts, Analysis tool (KNIME) | Data problems, technology problems |
| Modeling | 1 weeks | Analyst, some data analysis time, KNIME | Inability to find good model modal |
| Evaluation\Presentation | 1 week | All analysts | Inability to satisfy business |

## Data Understanding

We were provided with 4 SAS dataset tables namely:

**Accident**

The Accident data file includes crash data with each record corresponding to an incident of accident involving one or more vehicles.

**Distract**

The Distract data file identifies if each of the drivers involves in crash was distracted prior to crash.

**Vehicle**

The Vehicle data file includes in-transport motor vehicle data as well as driver and precast data.

**Person**

The Person data file includes all the motorist and non-motorists involves in an accident

We were also provided with (NASS)(GES) Analytical user’s manual. Thus, we could study the document and have a clear understanding of each of the variables in our dataset.

## DATA PREPROCESSING:

Data preparation was the most time consuming and difficult part of the entire project and we have invested large amount of time for data preprocessing. We have listed the process in which we have achieved data preprocessing.

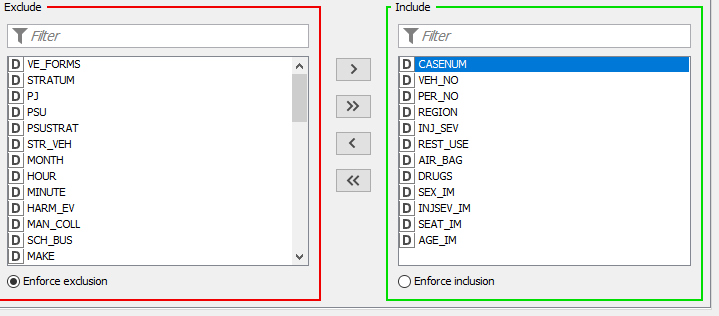
**Variable Selection:**

This was the vital part of data preprocessing. We have selected the variables after gathering relevant domain knowledge and brainstorming. Also, when we reiterated the process for improvement of the model, we have also considered the output of forward feature selection and backward variable elimination along with our perception.

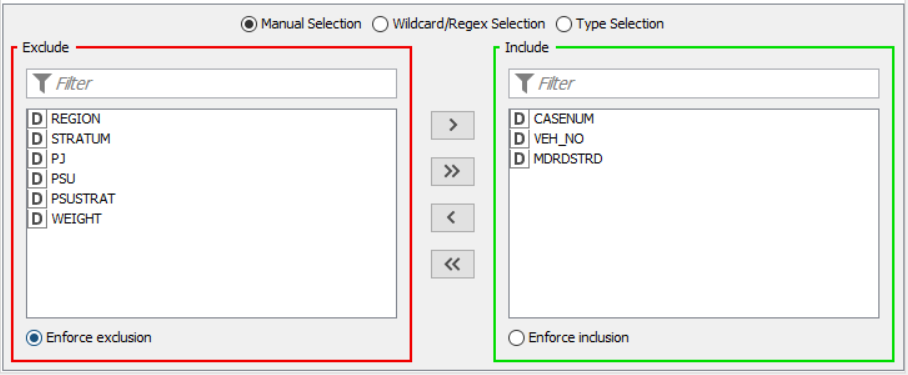
After variable selection, we have filtered out the necessary columns using column filter.

Following are the columns selected from each table:

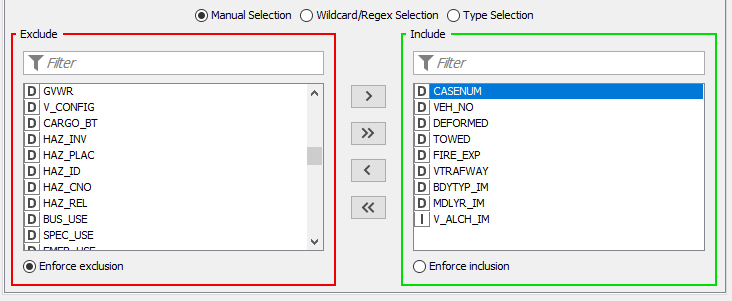
**Person Table:**



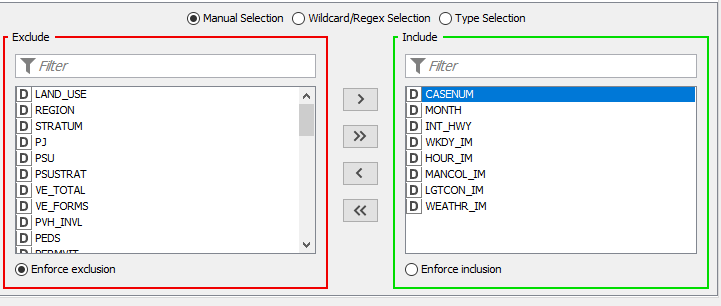
**Distract Table:**



**Vehicle Table:**



**Accident Table:**



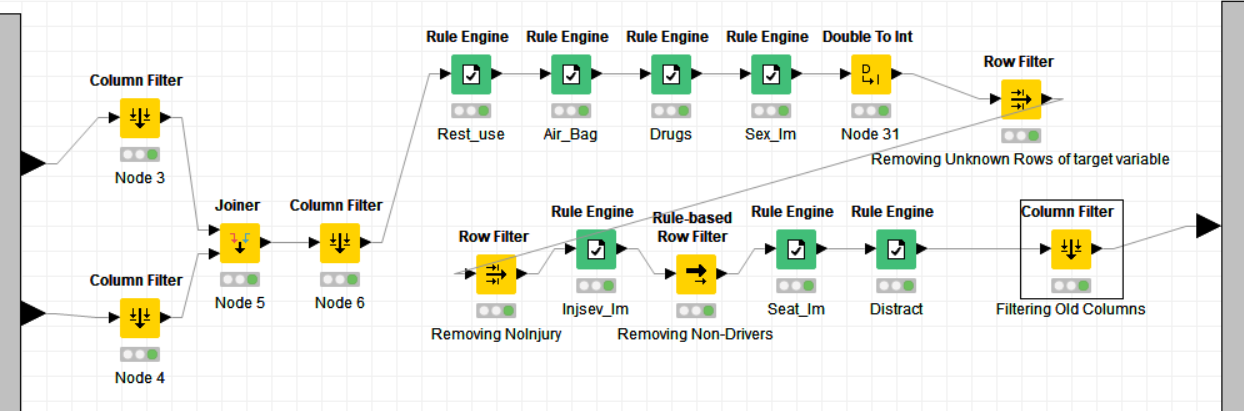
**Final Columns:**

The final 21 independent variables and the Injsev\_Im\_Binning dependent variables are as follows:

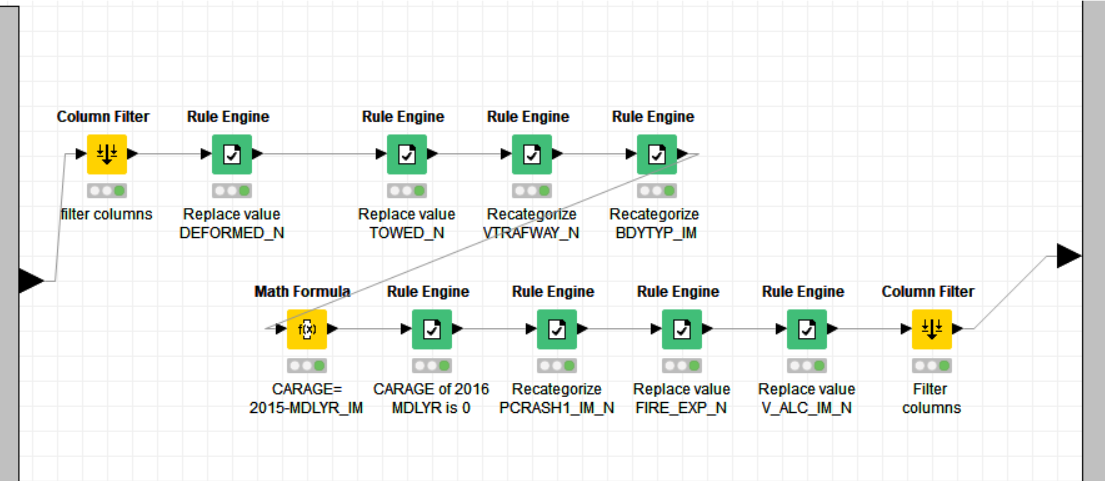
|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Variable Meaning** | **Attribute Type** | **Source Table** |
| REGION | Accidents in the four regions | Nominal | Person |
| AGE\_IM | Age of the drive | Numeric | Person |
| Rest\_use\_binning | Usage of safety restrictions | Nominal | Person |
| Air\_Bag\_Binning | Usage of airbags | Nominal | Person |
| Drugs\_Binning | Whether the drive takes drugs before the accident | Nominal | Person |
| Sex\_Im\_Binning | Sex of the driver | Nominal | Person |
| Distract\_Binning | Whether the driver is distracted before the accident | Nominal | Distract |
| DEFORMED\_N | Whether the vehicle is deformed or how deformed in crash | Nominal | Vehicle |
| TOWED\_N | Whether the vehicle is towed after crash | Nominal | Vehicle |
| VTRAFWAY\_N | Trafficway information | Nominal | Vehicle |
| BDYTYP\_IM\_N | Body types of the vehicles | Nominal | Vehicle |
| CARAGE | Car age | Numeric | Vehicle (derived) |
| FIRE\_EXP\_N | Whether fire exists | Nominal | Vehicle |
| V\_ALCH\_IM\_N | Whether the driver drink alcohol before the crash | Nominal | Vehicle |
| LGTCON\_IM\_B | Light conditions | Nominal | Accident |
| MANCOL\_IM\_B | Manners of collisions | Nominal | Accident |
| WKDY\_IM\_B | Which day in a week | Nominal | Accident |
| Weather\_IM\_B | Weather information | Nominal | Accident |
| Month\_Binning | Season information | Nominal | Accident |
| INT\_HWY\_B | Whether the crash happens in interstate highways | Nominal | Accident |
| Hour\_Binning | What time period the crash happens in a day | Nominal | Accident |
| Injsev\_Im\_Binning | The injury severity of the driver | Nominal | Person |

After variable selection, we have binned the categorical variables using Rule Engine Nodes. Also, we tried to limit the number of bins to 6.

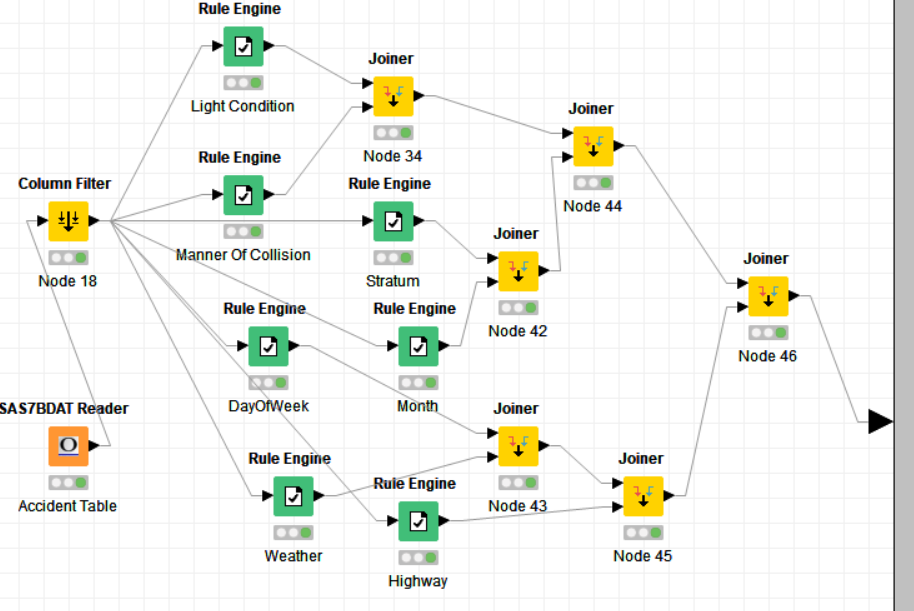
**Person+Distract:**



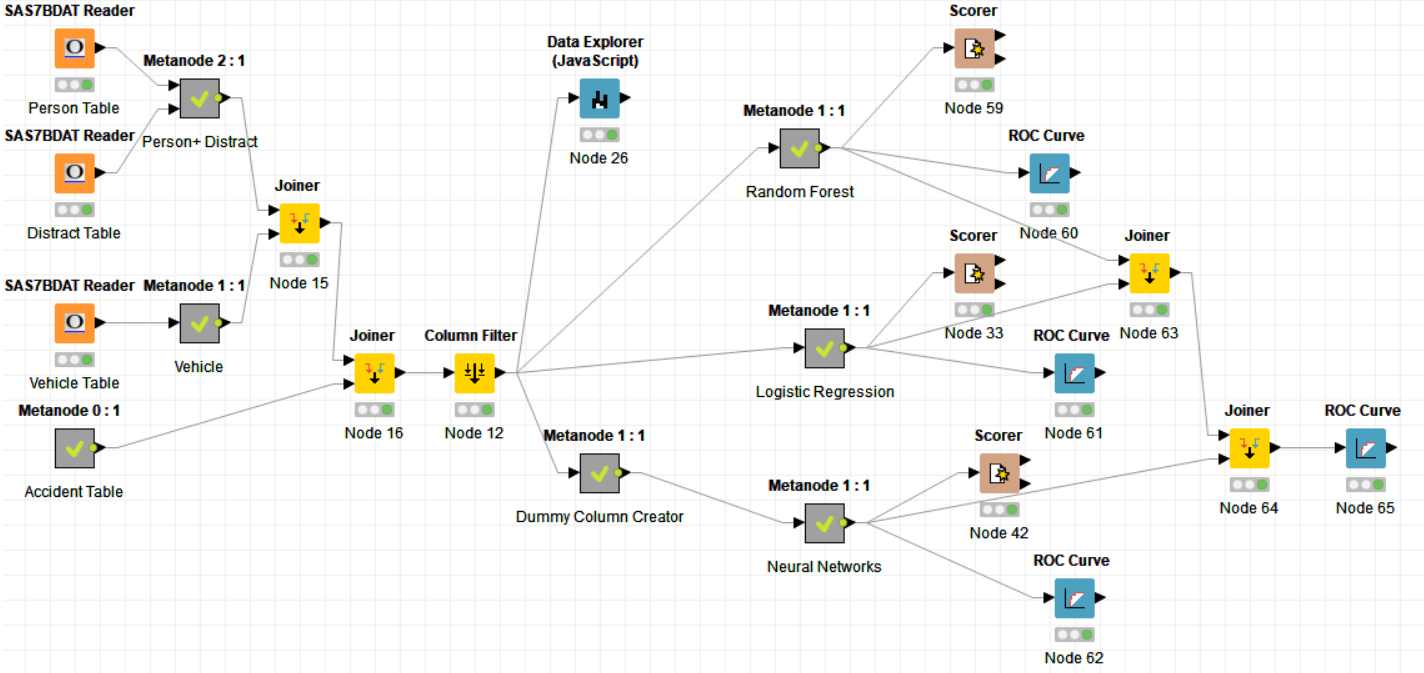
**Vehicle:**



**Accident:**



After binning variables, we have joined all the datasets.

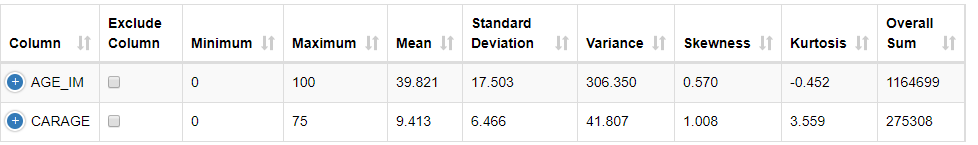


After joining the datasets, it was important to understand the data we have. For this we have relied upon the output of Data Explorer node. Following are the descriptive statistics of the variables we have considered for the models.

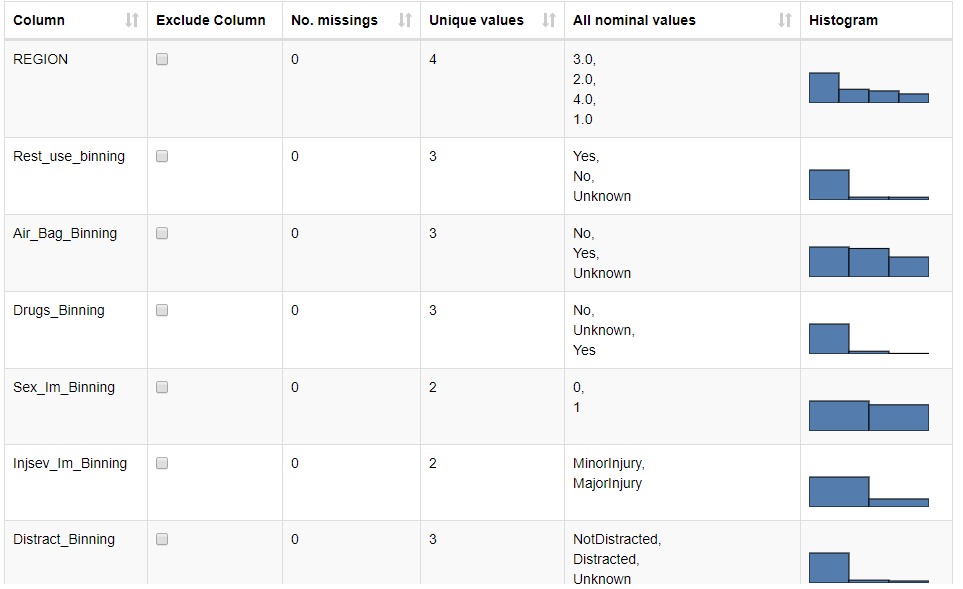
## Descriptive Statistics:

Target Variable: Injsev\_Im\_Binning

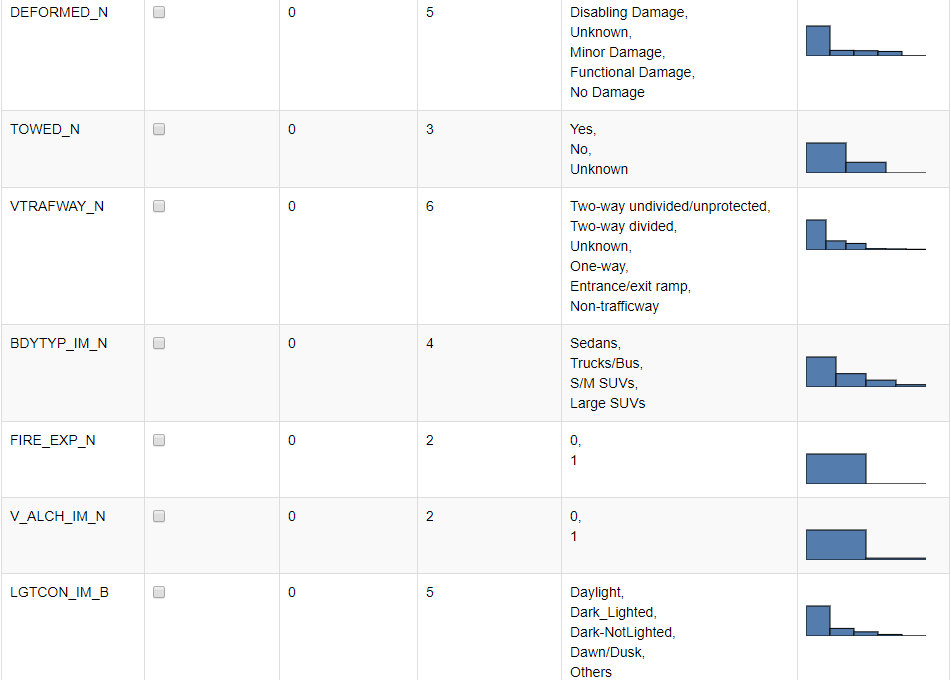
**Numeric Variables:**

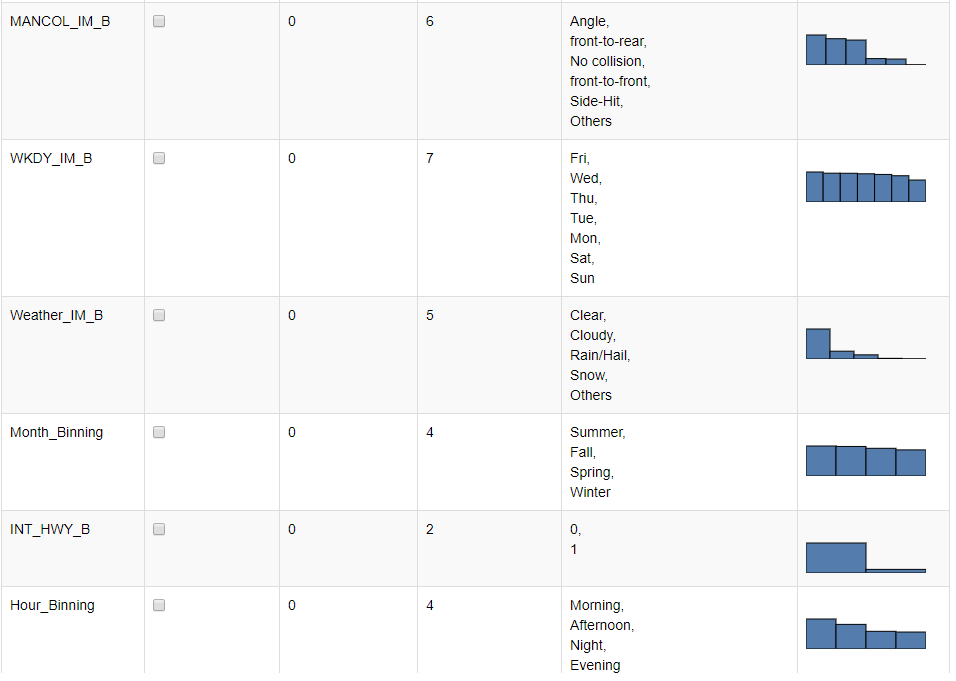


**Nominal Variables:**

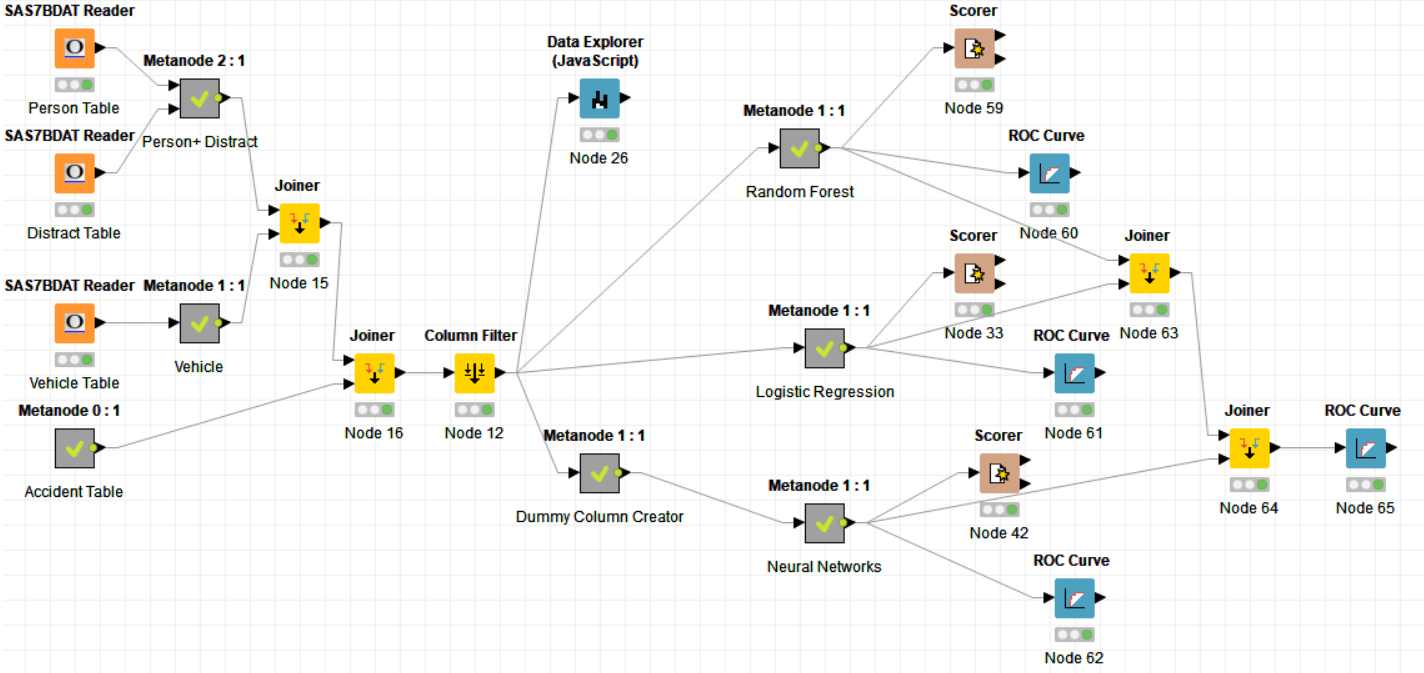


Here, you can observe that our target variable Injsev\_Im\_Binning, has a greater number of minor injury records – 23049 compared to major injury records- 6199.



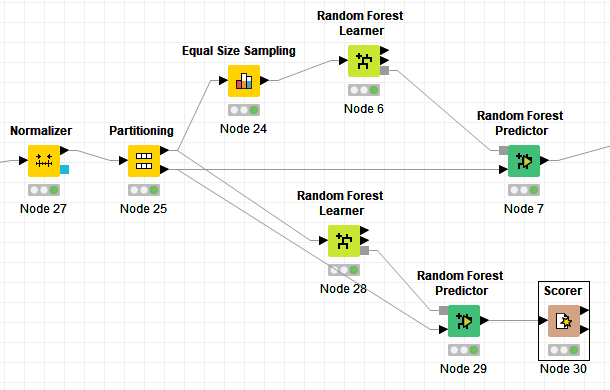


## MODEL BUILDING:



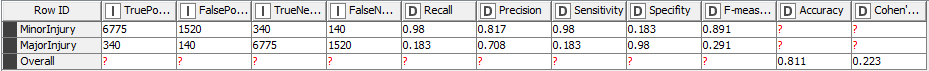
**Random Forest Model:**

In this model, we partitioned the dataset to two subsets, one training 70% and one testing 30%. We compared the performances between equal-size sampling method and no-sampling.

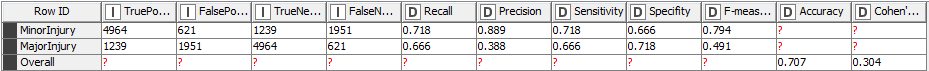


|  |  |
| --- | --- |
| Without Balancing: | With Balancing (Equal size sampling): |
|  |  |

Without Balancing:

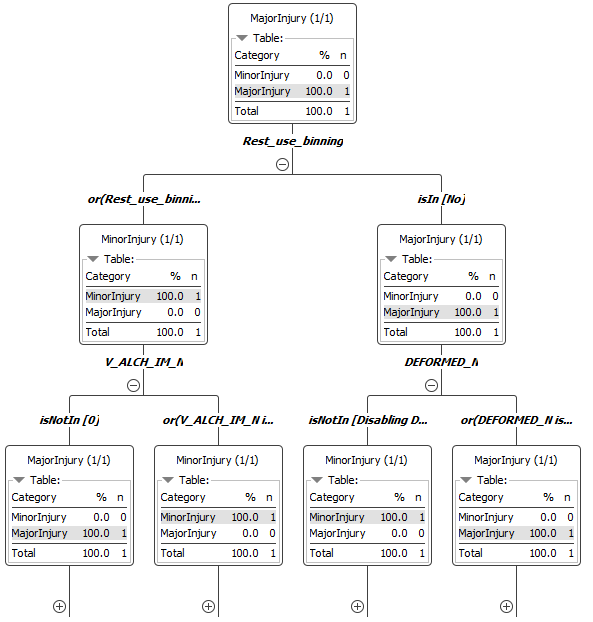


With Balancing:



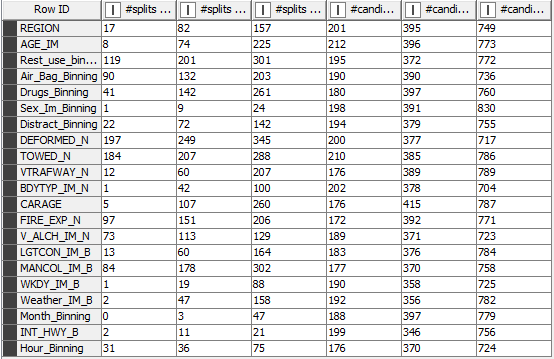
As we can see, with equal size sampling method, the predicting accuracy decreased from 81.1% to 70.7%, however, the sensitivity of MajorInjury increased from an unacceptable 18.3% to **66.6%**. Balancing helps!

Tree View (One of the 1000 trees generated):



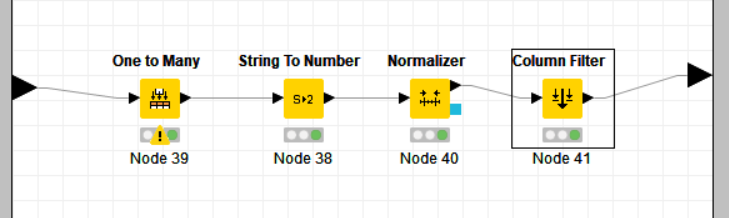
The first three root nodes are Rest\_use\_binning, V\_ALCH\_IM\_N and DEFORMED\_N.

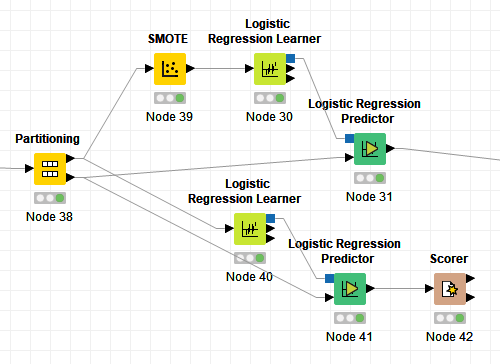
Attribute Statistics:



**Logistic Regression Model:**

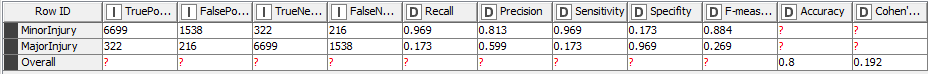
Dummy Variable Creation:





|  |  |
| --- | --- |
| Without Balancing: | With Balancing (SMOTE): |
|  |  |

Without Balancing:

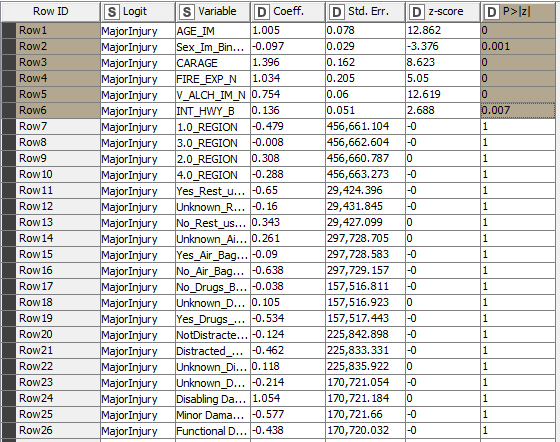


With Balancing:



As we can see, with SMOTE balancing method, the predicting accuracy decreased from 80 % to 68.7%, however, the sensitivity of MajorInjury increased from an unacceptable 17.3% to **66.8%**. Balancing helps again!

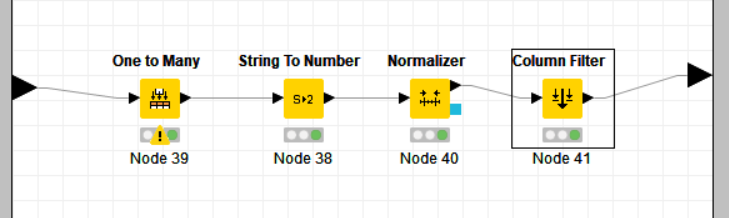
Significant Variables(highlighted):

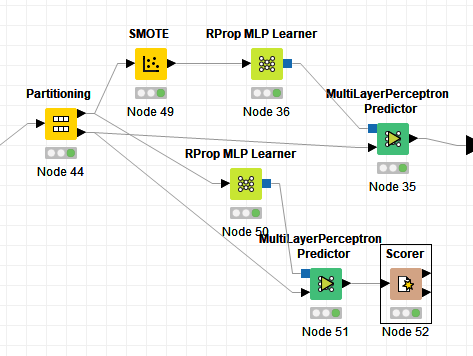


From the above Coefficients and Statistics table, we can conclude AGE\_IM, Sex\_Im\_Binning, CARAGE, FIRE\_EXP\_N, V\_ALCH\_IM\_N, and INT\_HWY\_B are statistically significant predictors of injury severity (p-values less than 0.05). Also, from the magnitudes and signs of coefficients, we can see older-aged male drivers with older-aged vehicles, drinking alcohol before involving crashes in interstate highways with fire occurrence are more likely to be severe injured.

**Neural Network Model:**

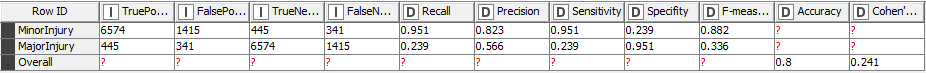
Dummy Variable Creation:



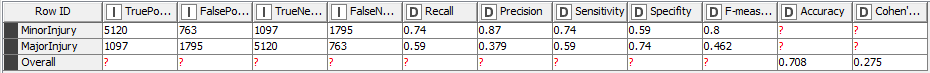


|  |  |
| --- | --- |
| Without Balancing: | With Balancing (SMOTE): |
|  |  |

Without Balancing:



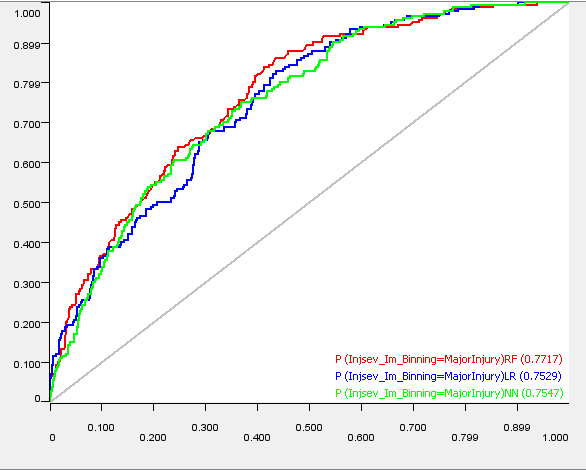
With Balancing:



As we can see, with SMOTE balancing method, the predicting accuracy decreased from 80% to 70.8%, however, the sensitivity of MajorInjury increased from an unacceptable 23.9% to **59%**. Balancing helps again!

## Model Evaluation/validation

**ROC Curve Comparisons:**



The above ROC curve on probability of predicting major injuries shows that the random forest method is the best one, beating logistic regression and artificial neural networks with slight better performances.

**Accuracy Statistics:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sl No | Model (balanced) | Accuracy | Majority Class Correctly classified(Specificity) | Minority class correctly classified (Sensitivity) |
| 1 | Random Forest | 70.7% | 71.8% | 66.6% |
| 2 | Logistic | 68.7% | 69.2% | 66.8% |
| 3 | Neural Network | 70.8% | 74% | 59% |

The above accuracy table shows that statistics of perfecting performances of the three balanced methods. The total predicting accuracies are around 70%, much the same. However, since we concern much more about major injuries, we should examine the sensitivity results, or the predicting accuracies of major injuries. Random forest and logistic regression outperformed artificial neural networks, both having a sensitivity of about 66.7%. Since the total accuracy of random forest is slightly higher than that of logistic regression, we propose random forest is the best method. This is also consistent with the ROC analysis.

Overall, the three models work well for the task to predict crash injury severity and we identify random forest as the best method.

## Deployment/ Suggestions:

This study reveals that certain factors are highly related with severe car crash injuries

* airbags and safety restrictions (e.g. safety belts) usages
* car deformed or not in accidents
* driver drinking / taking drugs or not
* manner of collision
* fire occurrence
* light conditions
* other factors

We suggest drivers:

* Always use safety restrictions (safety belts) properly
* Select cars with air bags and make sure they are not malfunctioned
* Never drink or take certain drugs (e.g. making people drowsy) before driving
* Be cautious about vehicles too old
* Male, old-aged drivers please drive more carefully☺

We suggest car makers:

* Invest and build more robust cars
* Include air bags and other safety installments
* Develop intelligent driving systems to identify risks and avoid accidents

We suggest authorities:

* Enforce strict surveillance on drunk driving and dangerous driving (e.g. safety belt unbuckled up)
* Invest in education on safe driving and accident emergency treatment.