Project Data Mining

Don’t get kicked!

Hamza Zafar,
Jehandad Kamal,
Moiz Zuberi

Table of Contents

Abstract: 3

Team: 3

Problem: 3

Analyzing Data: 3

Pre-processing: 6

- Removed Columns: 6
- Outlier Detection: 7
- Null Treatment: 7
- Misc: 7

Techniques applied: 9

- Decision Tree: 9
- Naïve Bayes: 11
- Neural Network – Using Rprop Learner: 12

Result: 13

1.0. Abstract:

This is a report for data mining project for academic semester of fall 2013. This consists of definition of problem at hand, the source the problem was extracted from and solutions taken to solve the problem. We have analyzed data, used a few preprocessing techniques to clean the data. Then the problem was modeled using cleaned data and Classifiers like Naïve Bayes and Decision Tree. The problem is competitions from kaggle with url http://www.kaggle.com/c/DontGetKicked.

Team:

- Abdul Moiz Zuberi
- Hamza Zafar
- Jehandad Kamal

2.0. Problem:

The problem is related to car buying in selling and our client i.e a car reseller want to know which cars can he buy and have a chance of making good profit from customers.

One of the biggest challenges of an auto dealership purchasing a used car at an auto auction is the risk of that the vehicle might have serious issues that prevent it from being sold to customers. The auto community calls these unfortunate purchases "kicks".

Kicked cars often result when there are tampered odometers, mechanical issues the dealer is not able to address, issues with getting the vehicle title from the seller, or some other unforeseen problem. Kick cars can be very costly to dealers after transportation cost, throw-away repair work, and market losses in reselling the vehicle.

Modellers who can figure out which cars have a higher risk of being kick can provide real value to dealerships trying to provide the best inventory selection possible to their customers.

The challenge of this competition is to predict if the car purchased at the Auction is a Kick (bad buy).

The rest of report, the data is analysed in Section 3. And the pre-processing is described in Section 4. We present our different result in Section 5.

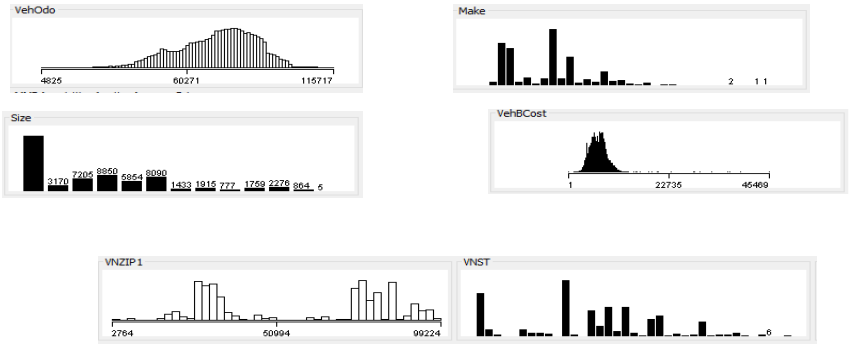
3.0. Analyzing Data:

There were 32 Columns of in our data each related to particular analyzed car. The size of data was 73015 records. Here a description of columns.

Field Name	Definition
RefID	Unique (sequential) number assigned to vehicles
IsBadBuy	Identifies if the kicked vehicle was an avoidable purchase
PurchDate	The Date the vehicle was Purchased at Auction
Auction	Auction provider at which the vehicle was purchased

VehYear	The manufacturer's year of the vehicle
VehicleAge	The Years elapsed since the manufacturer's year
Make	Vehicle Manufacturer
Model	Vehicle Model
Trim	Vehicle Trim Level
SubModel	Vehicle Submodel
Color	Vehicle Color
Transmission	Vehicles transmission type (Automatic, Manual)
WheelTypeID	The type id of the vehicle wheel
WheelType	The vehicle wheel type description (Alloy, Covers)
VehOdo	The vehicles odometer reading
Nationality	The Manufacturer's country
Size	The size category of the vehicle (Compact, SUV, etc.)
TopThreeAmericanName	Identifies if the manufacturer is one of the top three American manufacturers
MMRAcquisitionAuctionAveragePrice	Acquisition price for this vehicle in average condition at time of purchase
MMRAcquisitionAuctionCleanPrice	Acquisition price for this vehicle in the above Average condition at time of purchase
MMRAcquisitionRetailAveragePrice	Acquisition price for this vehicle in the retail market in average condition at time of purchase
MMRAcquisitonRetailCleanPrice	Acquisition price for this vehicle in the retail market in above average condition at time of purchase
MMRCurrentAuctionAveragePrice	Acquisition price for this vehicle in average condition as of current day
MMRCurrentAuctionCleanPrice	Acquisition price for this vehicle in the above condition as of current day
MMRCurrentRetailAveragePrice	Acquisition price for this vehicle in the retail market in average condition as of current day
MMRCurrentRetailCleanPrice	Acquisition price for this vehicle in the retail market in above average condition as of current day
PRIMEUNIT	Identifies if the vehicle would have a higher demand than a standard purchase
AcquisitionType	Identifies how the vehicle was aquired (Auction buy, trade in, etc)
AUCGUART	The level guarntee provided by auction for the vehicle (Green light - Guaranteed/arbitratable, Yellow Light – caution/issue, red light - sold as is)
KickDate	Date the vehicle was kicked back to the auction
BYRNO	Unique number assigned to the buyer that purchased the vehicle
VNZIP	Zipcode where the car was purchased
VNST	State where the the car was purchased
VehBCost	Acquisition cost paid for the vehicle at time of purchase
IsOnlineSale	Identifies if the vehicle was originally purchased online
WarrantyCost	Warranty price (term=36month and millage=36K)

The sample from the data is taken which shows the different behavior of data distribution. In data different columns were analyzed for missing values and distribution. As the result of analysis the data that was found to be skewed for both nominal and categorical data as shown in following figure.



4.0. Pre-processing:

4.1. Removed Columns:

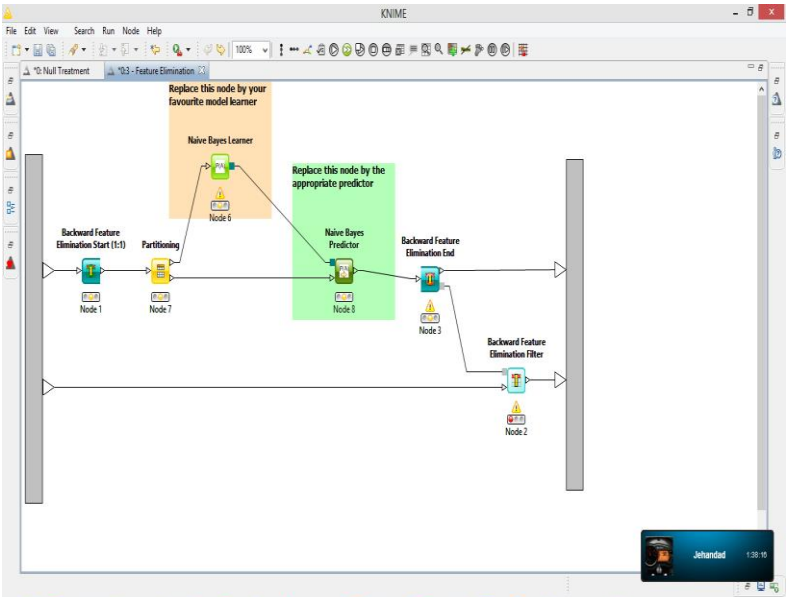
These columns were removed because they were id columns.

- RefID
- Buyer No
- WheelTypeId
- Purchase Date – Irrelevant

Using Backward Feature Elimination and testing we also removed these columns:

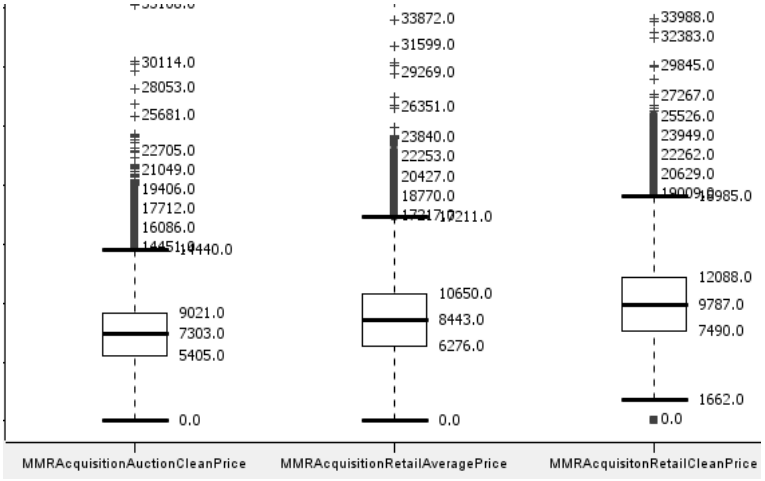
- MMRCurrentAuctionAveragePrice
- MMRAcquisitionAuctionAveragePrice
- Model

Here is KNIME Workflow used for feature elimination:

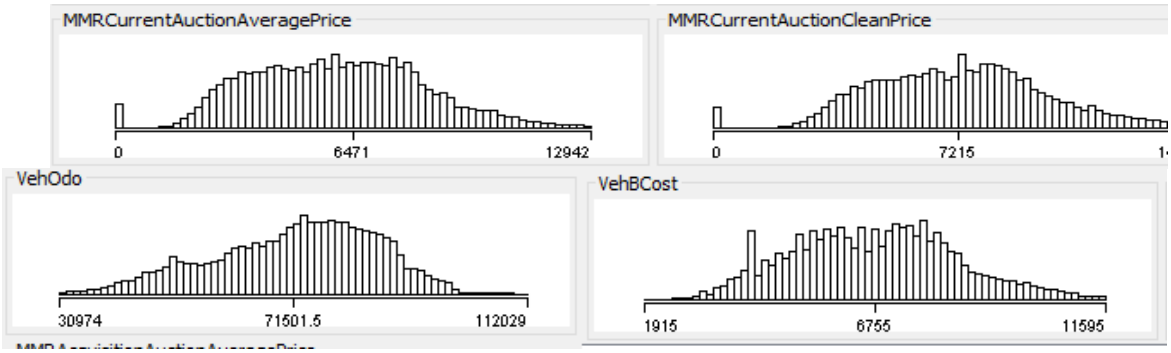


4.2. Outlier Detection:

All continuous data was fixed for skewed data using inter quartile ranges. Following four



columns showed excelled improvement from there normal skewed state after outlier detection.



4.3. Null Treatment:

All Columns were replaced by medians for nulls, in a few columns the replacement was not needed as null would itself represent a category.

4.4. Misc:

There were multiple date formats in columns

- Manufacturing date

They were converted into one format using excel.

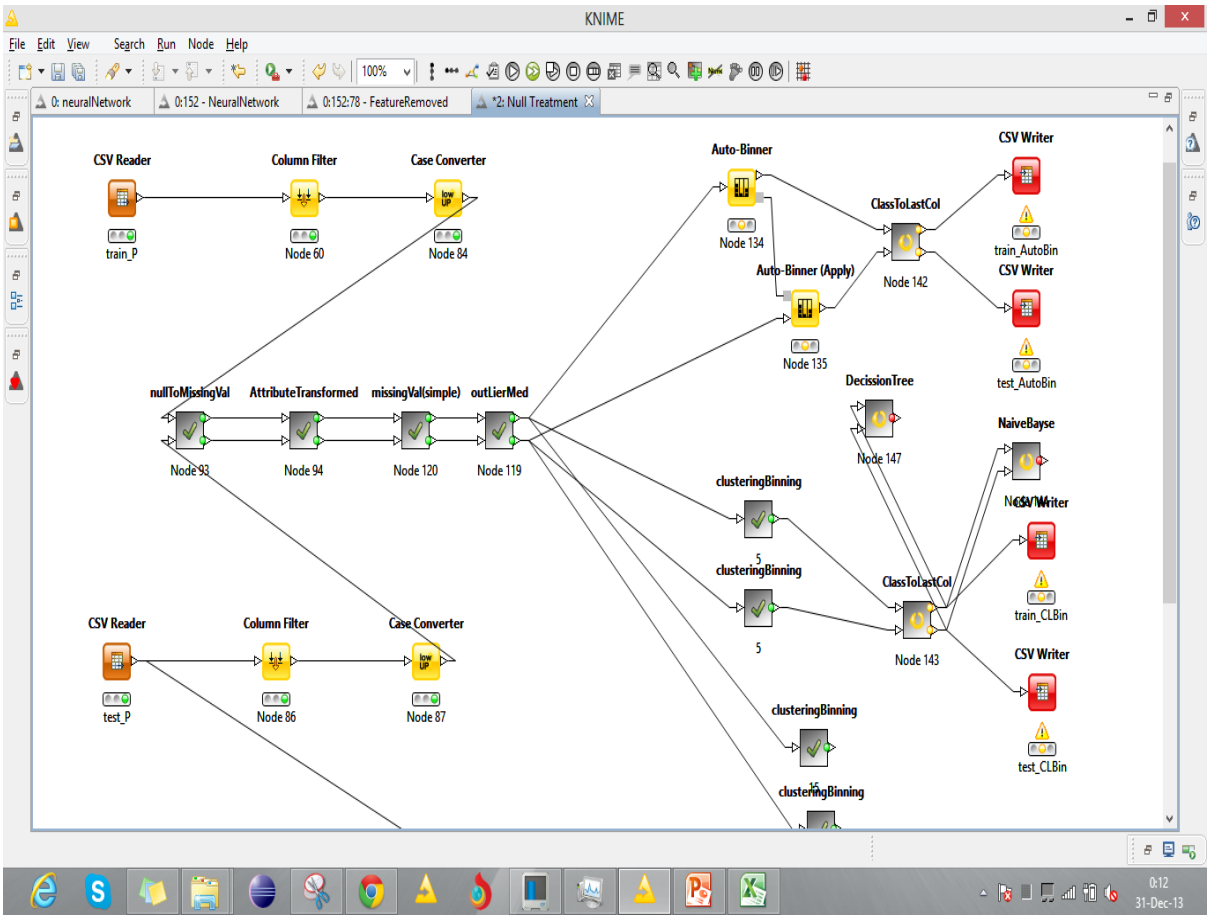
Also multiple columns were transformed into short hands to save memory.

COLUMN	VALUE	TRANSFORMED
Transmission	Auto	A
	Manual	M
Wheel type	Alloy	A
	Null	N
	Special	S
	Cover	C
Primeunit	Yes	Y
	No	N
	Null	Nl
Aucguart	Green	G
	Red	R
	Yellow	Y
	Null	N
topThreeName	chrysler	Ch
	Other	O
	Ford	F

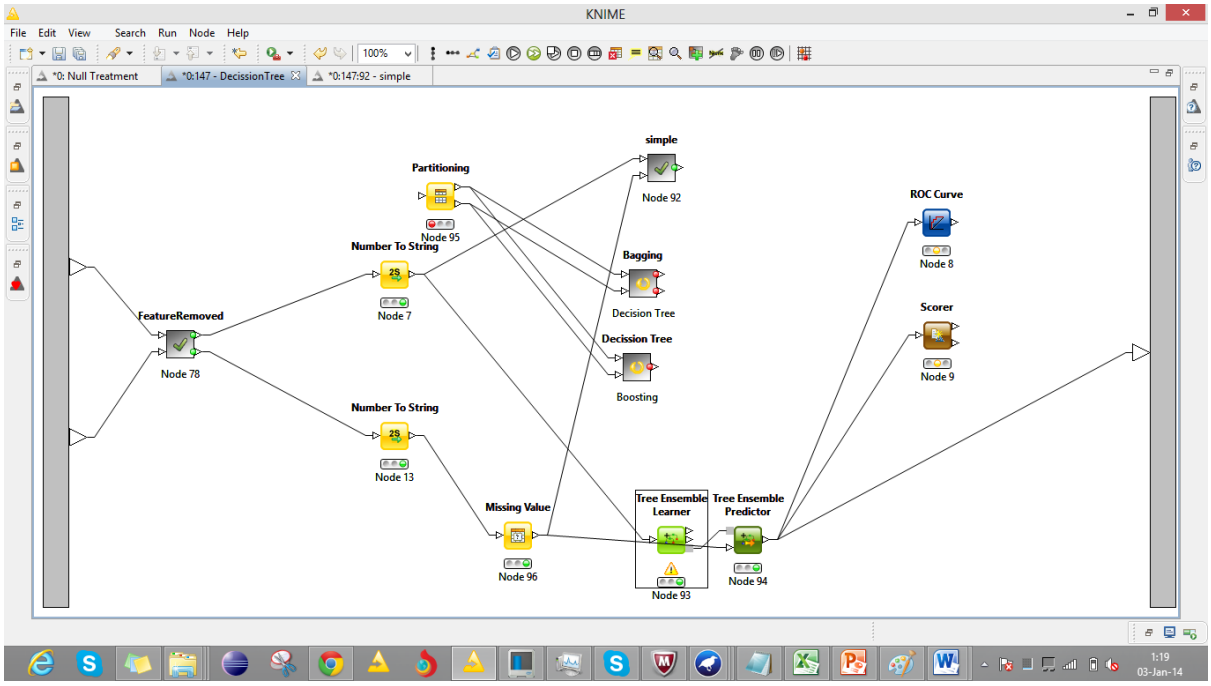
5.0. Techniques applied:

The following models provide of a walkthrough of the techniques used:

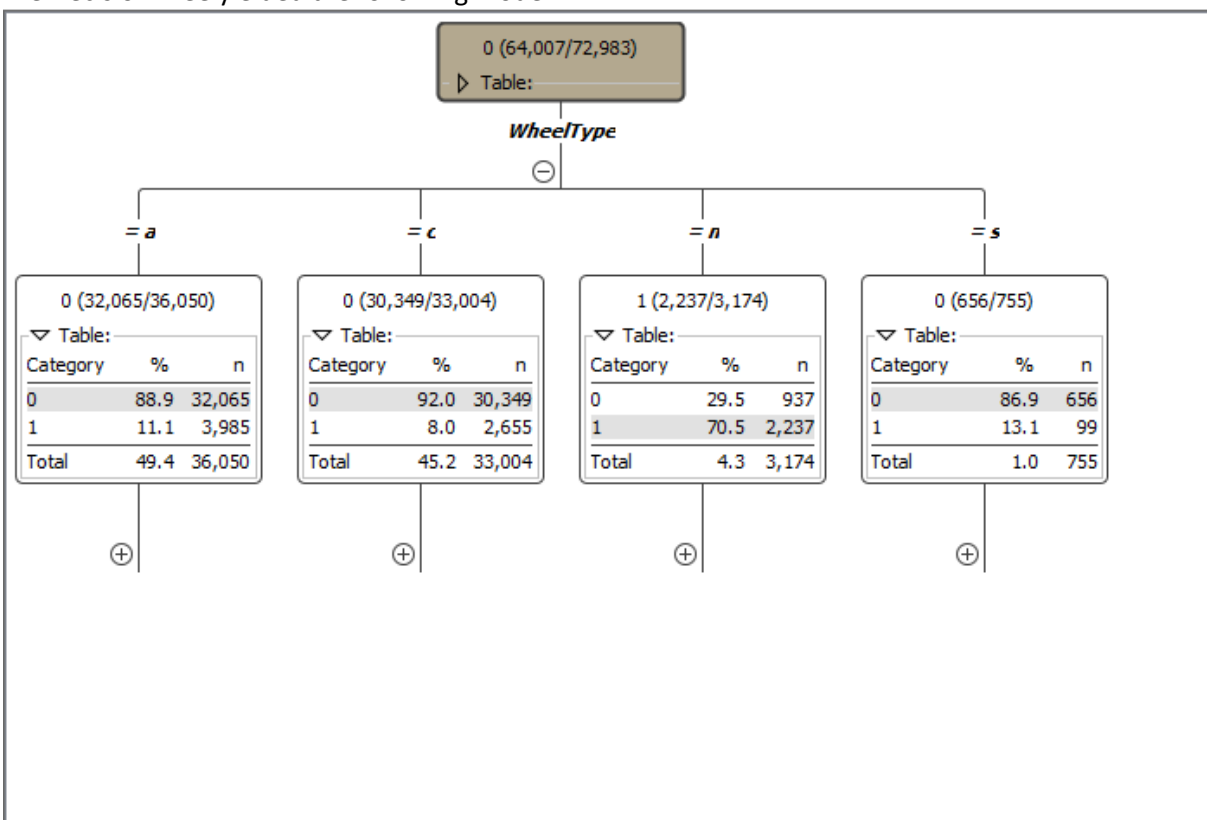
5.1. Overall Workflow:



5.2. Decision Tree:

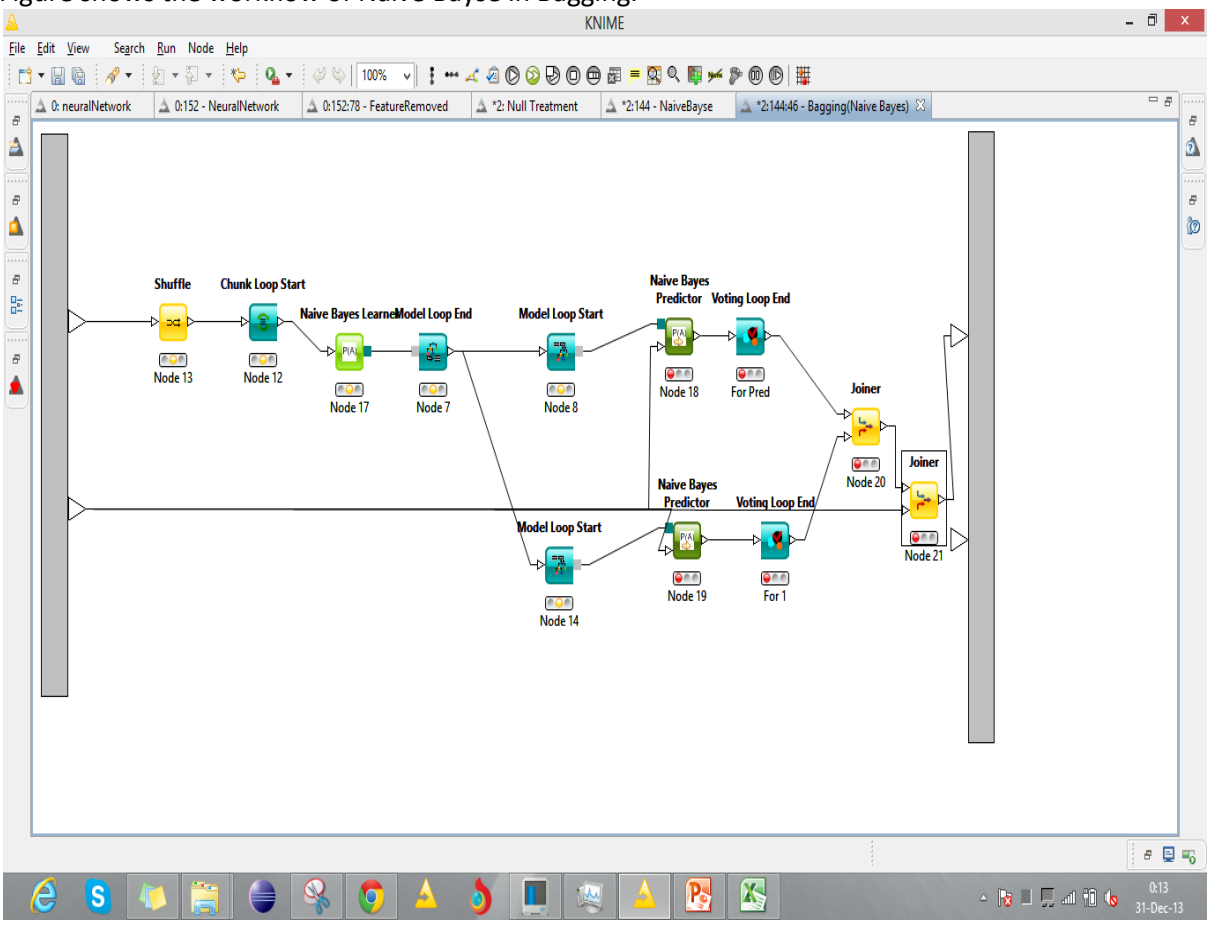


The Decision Tree yielded the following model:

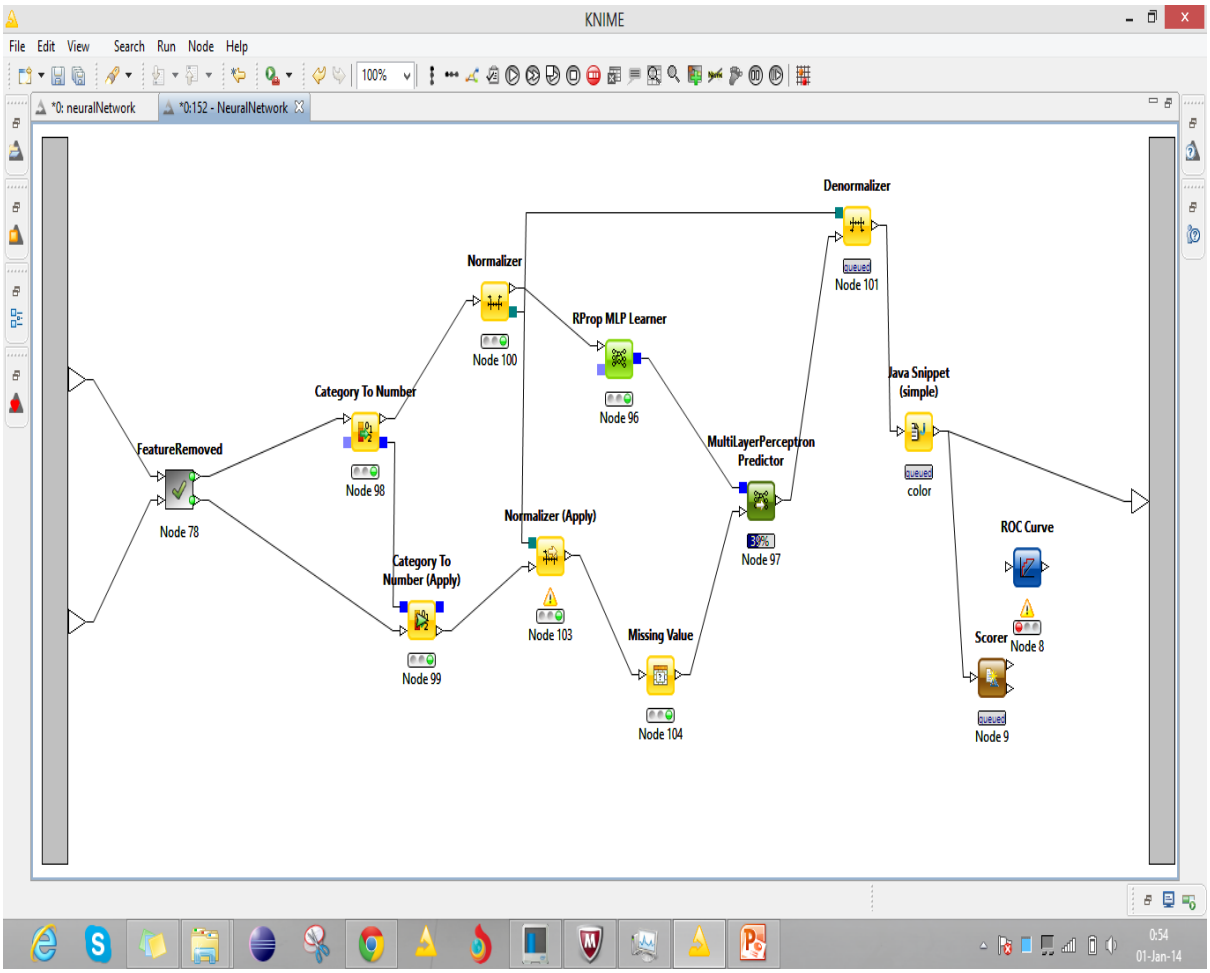


5.3. Naïve Bayes:

Figure shows the workflow of Naïve Bayse in Bagging.



5.4. Neural Network – Using Rprop Learner:

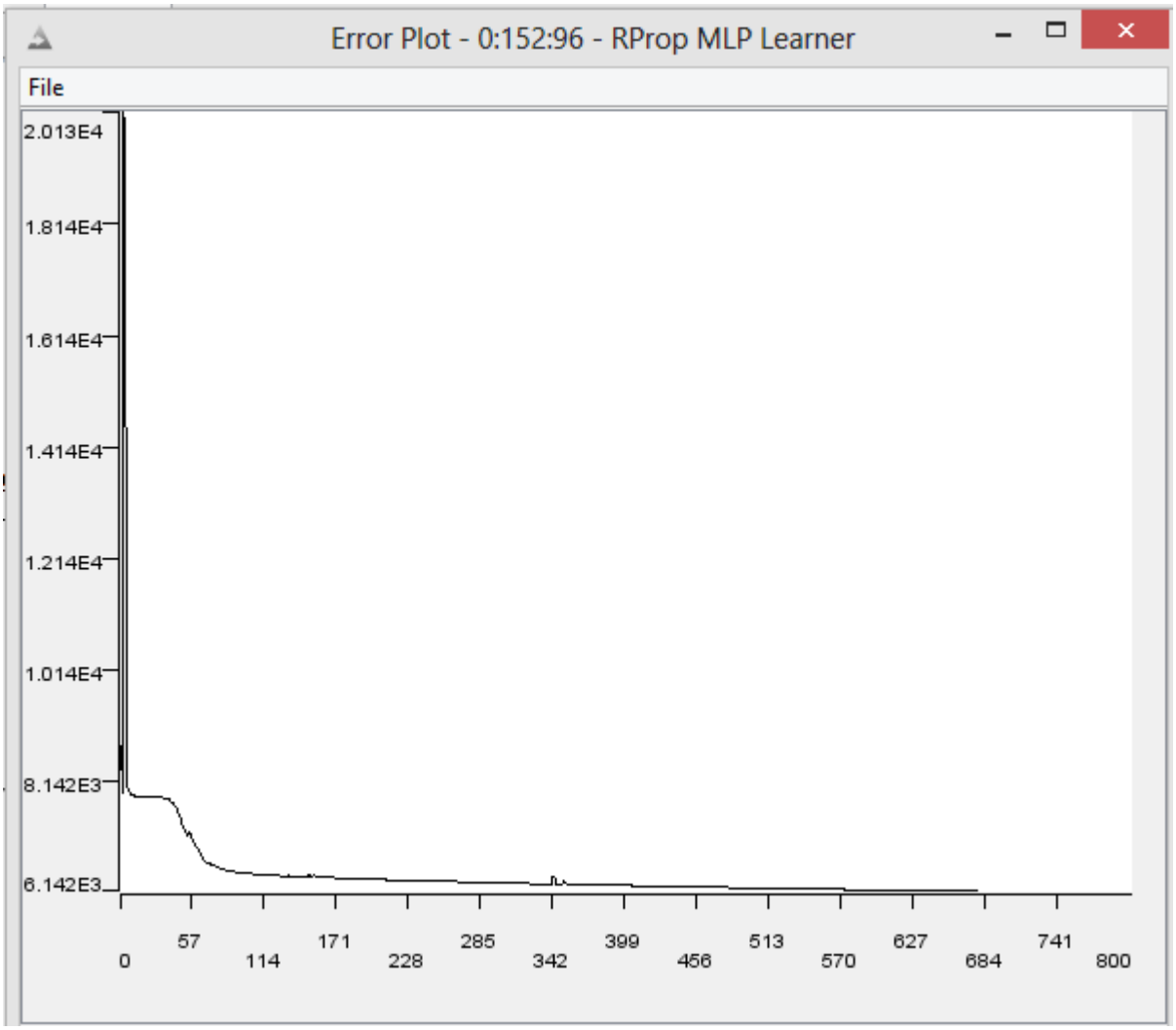


6.0. Result:

NAÏVE BAYSE	CLUSTER BINNING		AUTO-BINNER	
	ROC	FMEASURE	ROC	FMEASURE
Simple	0.696	0.318	0.693	0.317
Bagging	0.653	0.308	0.66	0.306
Boosting	0.694	0.315	0.693	0.317
DECISSION TREE				
Simple	0.597	0.318	0.603	0.322
Bagging	0.595	0.344	0.601	0.344
Boosting	0.607	0.316	0.624	0.317
Tree Ensembler	0.667	0.318	0.697	0.337

Neural Network

hidenLayer	no of Neuro	Iter	Gini
1	10	500	1.0043
1	10	1000	1.0076
1	10	2000	0.10187
10	25	400	0.10228
10	25	800	0.1053



7.0. Conclusion: