2020

MSIS 5663 – Data Warehousing

TERM PROJECT– SPRING 2020

SRIDHAR MOKKAPATI

A20224254

Table of Contents

Milestone 1 2
Named Calculations: 2
Measures 3
Hierarchy for Date Dimension 3
Second Hierarchy 4
MDX Queries: 4
Milestone 2 14
Create Statement for my Model(DMX Query): 14
Insert Statements DMX for Model: 14
Delete Statement DMX: 15
Alter Statements for creating different Models: 15
Decision Tree: 15
Neural Networks: 15
Clustering Model: 15
Model Comparison: 16
Lift Chart for all Models: 16
Classification Matrix: 20
Cross-Validation: 21
One Page Summary for Decision Tree Model: 25

Milestone 1

OLAP cube design and use (Due: Tuesday, April 21 in DropBox).

Create Cube with Dimensions Date, Accidents, Casualty, and Vehicles and created Named calculations for the descriptions as the Data contained numeric data. Also created some measure values mentioned below and queried the cube using MDX to draw some meaning full conclusions and values using various MDX functions.

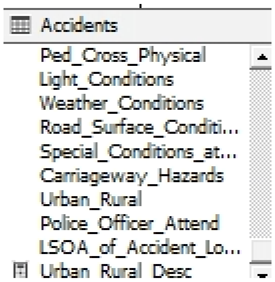
Named Calculations:

- Create *at least four* named calculations (as shown in Lecture 9) and use them in the dimensions.

- **Answer:**

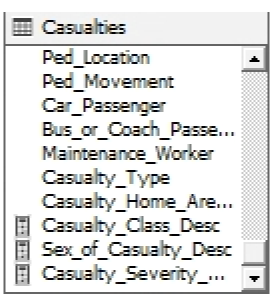
Added a Description column as follows for all the numeric values present in the tables.

Accidents Table:



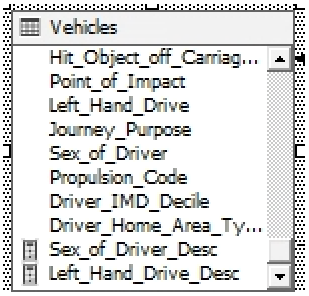
Urban_Rural_Desc

Casualties Table:



Casualty_Class_Desc, Sex_of_Casualty_Desc, Casualty_Severity_Desc

Vehicles Table:

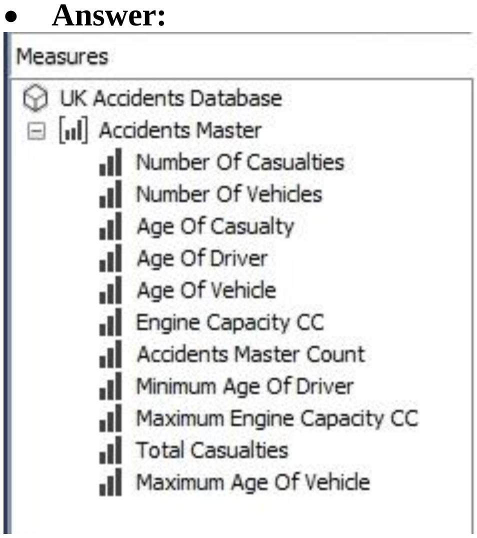


Sex_of_Driver_Desc, Left_Hand_Drive_Desc

Measures

- Create *at least* 2 new measures (as shown in Lecture 10) based on existing measures. Existing measures are: Number_of_Casualties, Number_of_Vehicles, Age_of_casualty, Age_of_driver, Age_of_vehicle_, Engine_Capacity_CC.

- **Answer:**



Minimum Age of Driver, Maximum Engine CC, Maximum Age of Vehicles, Total Casualties

The hierarchy for Date Dimension

- Create a hierarchy for the Date Dimension as similar to the AdventureWorksDW2012 analysis database as possible.

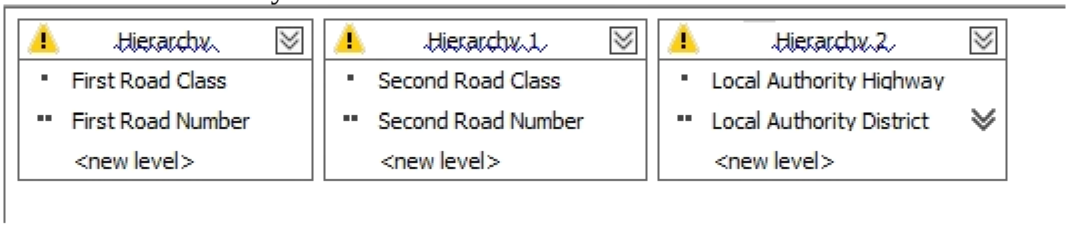
- **Answer:**

Second Hierarchy

- ○ Create a *meaningful hierarchy in at least one other dimension*.

- Answer:

  Created Hierarchy for Accidents Dimension as follows:



MDX Queries:

- ○ *Design, Execute and Report* **at least 10** MDX queries highlighting use of the OLAP cube you have designed. Your queries should highlight the use of **10 different MDX functions**.

1. Top Count of Causalities per year by the minimum aged Drivers.

```
SELECT
TOPCOUNT
(
{[Measures].[Number Of Casualties]}
   , 10
   , [Measures].[Minimum Age Of Driver]
)
ON COLUMNS,
[Date].[DateTime].[Year]
ON ROWS
FROM [UK Accidents Database]  ;
```

The output gives us the count of the top 10 causalities per year caused by the minimum aged drivers.

Output:

Messages Results

	Number Of Casualties
2005	1060832
2006	1004696
2007	993009
2008	904923
2009	857946
2010	814998
2011	932853
2012	753827
2013	1016354
2014	773049
2015	762726

2. Accidents by Year for Cambridge and Manchester District

```
SELECT [Measures].[Number Of Casualties] ON 0,
UNION(
{
[Accidents].[Local Authority District Desc].&[Cambridge] * [Date].[DateTime].[Year]
}
,
{
[Accidents].[Local Authority District Desc].&[Manchester] * [Date].[DateTime].[Year]
}
)
ON 1
FROM [UK Accidents Database]
```

Output:

		Number Of Casualties
Cambridge	2005	1510
Cambridge	2006	1579
Cambridge	2007	1323
Cambridge	2008	1399
Cambridge	2009	1275
Cambridge	2010	1399
Cambridge	2011	1078
Cambridge	2012	940
Cambridge	2013	858
Cambridge	2014	1040
Cambridge	2015	863
Manchester	2005	11979
Manchester	2006	11707
Manchester	2007	11309
Manchester	2008	9490
Manchester	2009	9481
Manchester	2010	7776
Manchester	2011	8150
Manchester	2012	5822
Manchester	2013	5741
Manchester	2014	7050
Manchester	2015	3599

3. Mean Casualties per accident in the three cities calculated as casualties/Accident.

```
WITH
SET [TOP] AS
'{
[Accidents].[Local Authority District Desc].&[Cambridge],
[Accidents].[Local Authority District Desc].&[Manchester],
[Accidents].[Local Authority District Desc].&[Nottingham]
}'

MEMBER [Measures].[Mean No of Casualties Per Accident] AS
([Measures].[Number Of Casualties]/[Measures].[Accidents Master Count])
SELECT {
[Measures].[Mean No of Casualties Per Accident]
} ON 0,
[TOP]
ON 1
FROM [UK Accidents Database]
```

Output:

	Mean No of Casualties Per Accident
Cambridge	1.35679214402619
Manchester	2.10177536397243
Nottingham	1.78668531033947

4. Total Number of accidents per year occurred where gender of the casualty is Female.

```
SELECT
      {[Measures].[Accidents Master Count]} ON COLUMNS,
FILTER
(
       {[Date].[Year].[Year]},  [Casualties].[Sex Of Casualty Desc].&[Female]

) ON ROWS
FROM [UK Accidents Database]
```

Output:

	Accidents Master Count
2005	525546
2006	500310
2007	480505
2008	444363
2009	427676
2010	401198
2011	395395
2012	377415
2013	358094
2014	377091
2015	363266

5. Bottom 10 Districts based on number of casualties

```
SELECT
      {[Measures].[Number Of Casualties]} ON COLUMNS, NON EMPTY
BOTTOMCOUNT
(
       {[Accidents].[Local Authority District Desc].MEMBERS}
```

```
        , 10
        , [Measures].[Measures].[Number Of Casualties]
) ON ROWS
FROM [UK Accidents Database]
```

Output:

Messages | Results

	Number Of Casualties
Orkney Islands	1227
London Airport (Heathrow)	1456
Western Isles	1726
Berwick-upon-Tweed	1784
Shetland Islands	1859
Teesdale	2086
Oswestry	2221
South Shropshire	2733
Alnwick	2795

6. Top 10 Districts based on the no of casualties.

```
SELECT
      {[Measures].[Number Of Casualties]} ON COLUMNS, NON EMPTY
TOPCOUNT
(
       {[Accidents].[Local Authority District Desc].children}
        , 10
        , [Measures].[Measures].[Number Of Casualties]
) ON ROWS
FROM [UK Accidents Database]
```

Output:

Messages | Results

	Number Of Casualties
Swale	353254
Birmingham	182032
Leeds	134932
Liverpool	111562
Taunton Deane	105659
Bradford	105236
Manchester	92104
Kirklees	79005
Salford	78091
Sheffield	77495

7. Top sum for the vehicles type that are involved in the accidents that are having casualty severity as fatal.

```
SELECT
    [Measures].[Accidents Master Count] ON COLUMNS,
TOPSUM
(
[Vehicles].[Vehicle Type Desc].children, 100000,[Measures].[Accidents Master Count]
) ON ROWS
FROM [UK Accidents Database]

 WHERE
[Casualties].[Casualty Severity Desc].&[Fatal]
```

Output:

Messages Results

	Accidents Master Count
Car	28804
Motorcycle over 500cc	4141
Goods 7.5 tonnes mgw and over	3566
Van / Goods 3.5 tonnes mgw or under	2423
Pedal cycle	1475
Bus or coach (17 or more pass seats)	982
Motorcycle 125cc and under	722
Goods over 3.5t. and under 7.5t	661
Motorcycle over 125cc and up to 500cc	561
Taxi/Private hire car	540
Other vehicle	509
Agricultural vehicle	291
Minibus (8 - 16 passenger seats)	198
Motorcycle 50cc and under	166
Ridden horse	25
Mobility scooter	23
Goods vehicle - unknown weight	19

8. Casualty Rank for Vehicle Type based on Severity

```
WITH
SET [Top 5 VehicleType] AS
TopCount ([Vehicles].[Vehicle Type Desc].children, 5,
[Measures].[Accidents Master Count])

MEMBER [Measures].[Vehicle Type Casualty Rank] AS
RANK ([Vehicles].[Vehicle Type Desc].CurrentMember,
[Top 5 VehicleType]
)

SELECT
{[Measures].[Accidents Master Count] ,
       [Measures].[Vehicle Type Casualty Rank]  }
ON 0,
[Top 5 VehicleType]
ON 1
FROM
[UK Accidents Database]
```

Output:

	Accidents Master Count	Vehicle Type Casualty Rank
Car	3553957	1
Van / Goods 3.5 tonnes mgw or under	217491	2
Pedal cycle	213241	3
Bus or coach (17 or more pass seats)	115891	4
Motorcycle over 500cc	112021	5

9. Correlation

```
WITH
      MEMBER [Measures].[CorrCoef] AS
      CORRELATION([Date].[Year].[Year].Members,
       [Measures].[Accidents Master Count],
       [Measures].[Number Of Vehicles])
SELECT
      {[Measures].[CorrCoef]}
ON COLUMNS
FROM [UK Accidents Database]
```

Output:

CorrCoef
0.760751028236562

10. Total accidents based on Severity of Casualty and each type of light conditions

```
SELECT
 [Measures].[Accidents Master Count]  ON COLUMNS,
 NON EMPTY
 ({[Accidents].[Light Conditions Desc].children},
 {[Accidents].[Accident Severity Desc].[Accident Severity Desc]})  ON ROWS
FROM [UK Accidents Database];
```

Output:

		Accidents Master Count
Darkness - lighting unknown	Fatal	800
Darkness - lighting unknown	Serious	5907
Darkness - lighting unknown	Slight	39369
Darkness - lights lit	Fatal	14809
Darkness - lights lit	Serious	126474
Darkness - lights lit	Slight	739758
Darkness - lights unlit	Fatal	564
Darkness - lights unlit	Serious	2971
Darkness - lights unlit	Slight	17005
Darkness - no lighting	Fatal	17247
Darkness - no lighting	Serious	62261
Darkness - no lighting	Slight	185063
Daylight	Fatal	57087
Daylight	Serious	451525
Daylight	Slight	2930019

11. Order of Accident Total based on vehicle type and remove unknown columns.

```
SELECT
{
[Measures].[Accidents Master Count],
[Measures].[Number Of Vehicles]
}
ON COLUMNS,
ORDER(
EXCEPT(
[Vehicles].[Vehicle Type Desc].[Vehicle Type Desc].members,
[Vehicles].[Vehicle Type Desc].[All].UNKNOWNMEMBER
),
[Measures].[Accidents Master Count], DESC
)
ON ROWS
FROM [UK Accidents Database]
```

Output:

Messages Results

	Accidents Master Count	Number Of Vehicles
Car	3553957	8525184
Van / Goods 3.5 tonnes mgw or under	217491	604276
Pedal cycle	213241	432191
Bus or coach (17 or more pass seats)	115891	202955
Motorcycle over 500cc	112021	232705
Goods 7.5 tonnes mgw and over	96387	261910
Motorcycle 125cc and under	86733	163172
Taxi/Private hire car	81693	170309
Motorcycle 50cc and under	42909	81655
Other vehicle	34143	75282
Goods over 3.5t. and under 7.5t	33962	83281
Motorcycle over 125cc and up to 500cc	31685	60786
Minibus (8 - 16 passenger seats)	16165	35509
Agricultural vehicle	9169	20590
Ridden horse	1739	4375
Goods vehicle - unknown weight	1314	3261
Mobility scooter	707	1258
Motorcycle - unknown cc	636	1279
Data missing or out of range	619	1287
Tram	367	642
Electric motorcycle	30	52

Milestone 2

- Use at least three different data mining techniques based on the Microsoft data mining algorithms available in Analysis Services on the data (you can use the cube or the relational tables in **UK_Accidents_Database** directly).

Answer:

I performed the analysis by creating Decision Tree, Neural Network, and Clustering Models and used DimAccidents Table which I renamed it as Accidents in my Deliverable- I to create Accidents Dimension.

Create Statement for my Model(DMX Query):

```
CREATE MINING STRUCTURE [Accident Severity Models DMX]
(

[Weather Conditions] TEXT DISCRETE,
[Urban Rural] LONG DISCRETE,
[Speed Limit] TEXT DISCRETE,
[Road Type] LONG DISCRETE,
[Road Surface Conditions] TEXT DISCRETE,
[Local Authority Highway] TEXT DISCRETE,
[Local Authority District] LONG DISCRETE,
[Light Conditions] TEXT DISCRETE,
[Accident Severity] LONG DISCRETE,
[Accident Index] TEXT KEY
)
WITH HOLDOUT (30 PERCENT or 1000 CASES)
```

Insert Statements DMX for Model:

```
INSERT INTO MINING STRUCTURE [Accident Severity Models DMX]
(
[Weather Conditions],
[Urban Rural],
[Speed Limit],
[Road Type],
[Local Authority Highway] ,
[Local Authority District],
[Light Conditions],
[Accident Severity],
[Accident Index]
)
OPENQUERY([UK Accidents Database],
'SELECT
A.Weather_Conditions,
A.Urban_Rural,
A.Speed_Limit,
A.Road_Type,
A.Local_Authority_Highway,
A.Local_Authority_District,
A.Light_Conditions,
A.Accident_Severity,
A.Accident_Index
   FROM  dbo.DimAccidents A'
)
```

Delete Statement DMX:

```
DELETE FROM [Accident Severity Models DMX]
```

Alter Statements for creating different Models:

Decision Tree:

```
ALTER MINING STRUCTURE [Accident Severity Models DMX]
ADD MINING MODEL [Decision Tree Model DMX]
(
[Weather Conditions],
[Urban Rural],
[Speed Limit],
[Road Type],
[Local Authority Highway],
[Local Authority District],
[Light Conditions],
[Accident Severity] PREDICT,
[Accident Index]
) USING Microsoft_Decision_Trees
WITH DRILLTHROUGH
```

Neural Networks:

```
ALTER MINING STRUCTURE [Accident Severity Models DMX]
ADD MINING MODEL [Neural Network Model DMX]
(
[Weather Conditions],
[Urban Rural],
[Speed Limit],
[Road Type],
[Local Authority Highway],
[Local Authority District],
[Light Conditions],
[Accident Severity] PREDICT,
[Accident Index]
)
```

Clustering Model:

```
ALTER MINING STRUCTURE [Accident Severity Models DMX]
ADD MINING MODEL [Neural Network Model DMX]
(
[Weather Conditions],
[Urban Rural],
[Speed Limit],
[Road Type],
[Local Authority Highway],
[Local Authority District],
[Light Conditions],
[Accident Severity] PREDICT,
[Accident Index]
) USING Microsoft_Clustering
WITH DRILLTHROUGH
```

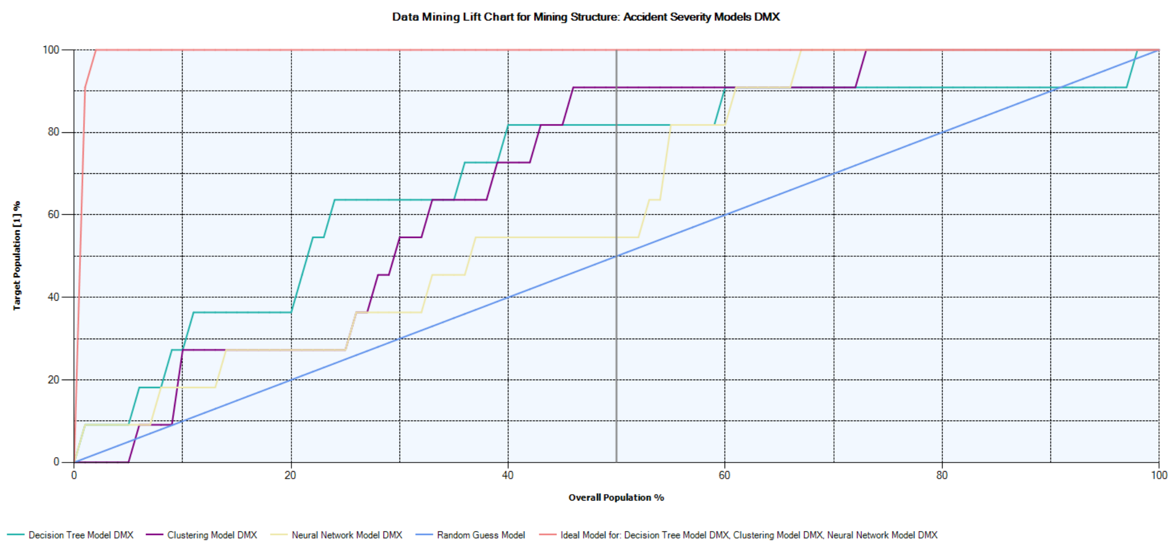
- Compare the performance of the three algorithms using the various metrics, conclude which is the best model.

Answer:

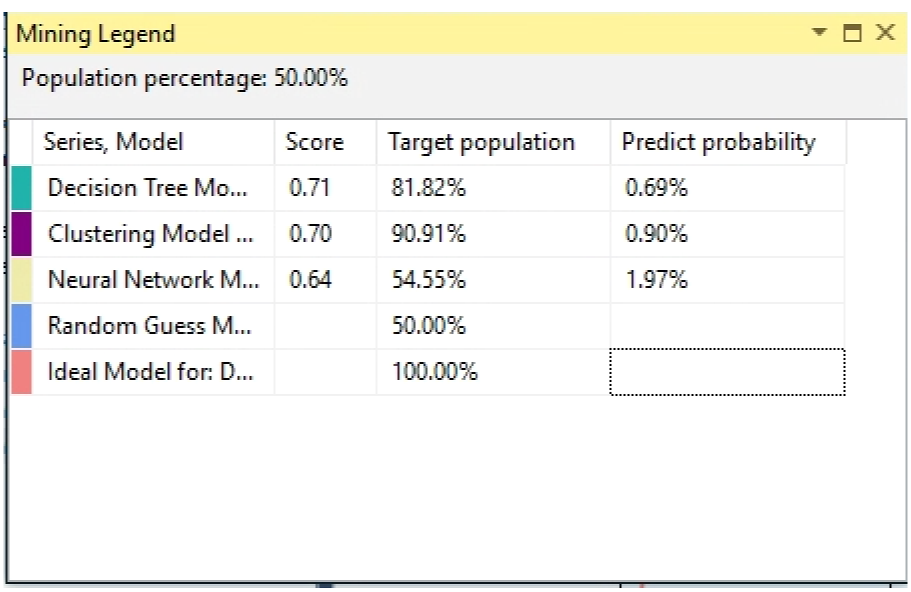
Model Comparison:

Lift Chart for all Models:

For target state as 1 for the predict variable Accident_Severity in the accident table.



Lift Score:

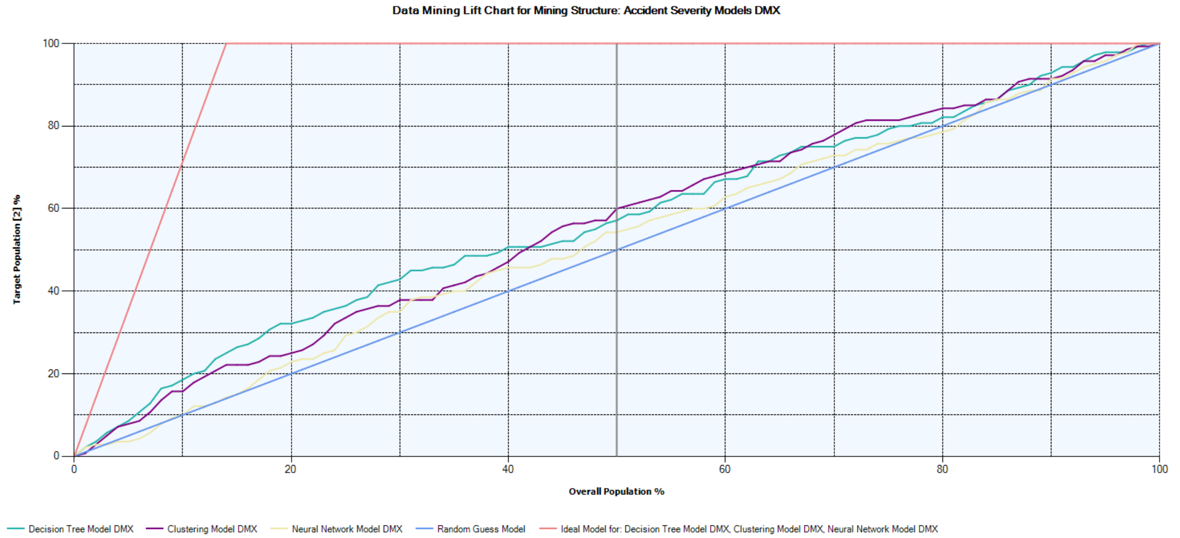


Mining Legend

Population percentage: 50.00%

Series, Model	Score	Target population	Predict probability
Decision Tree Mo...	0.71	81.82%	0.69%
Clustering Model ...	0.70	90.91%	0.90%
Neural Network M...	0.64	54.55%	1.97%
Random Guess M...		50.00%	
Ideal Model for: D...		100.00%	

For target state as 2 for the predict variable Accident_Severity in the accident table.



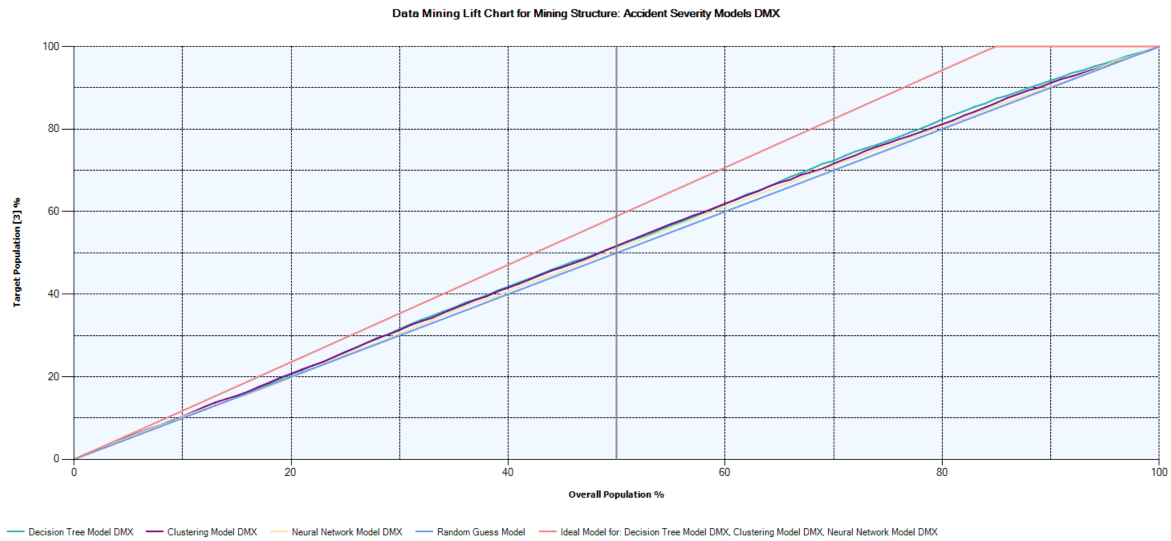
Lift Score:

Mining Legend

Population percentage: 50.00%

Series, Model	Score	Target population	Predict probability
Decision Tree Mo...	0.61	57.14%	11.93%
Clustering Model ...	0.60	60.00%	12.63%
Neural Network M...	0.56	54.29%	11.28%
Random Guess M...		50.00%	
Ideal Model for: D...		100.00%	

For target state as 3 for the predict variable Accident_Severity in the accident table.



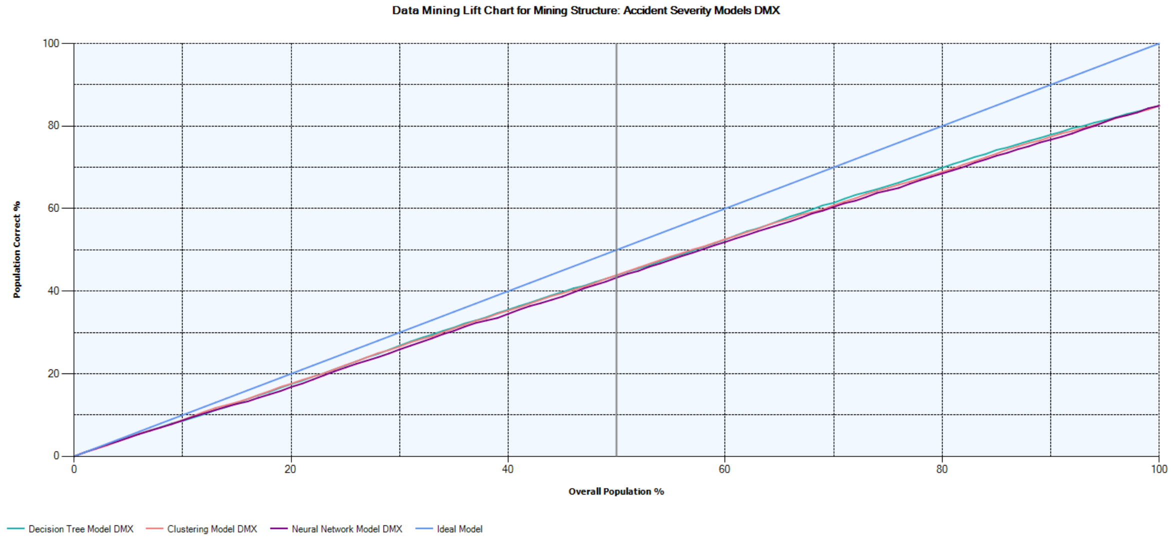
Lift Score:

Mining Legend

Population percentage: 50.00%

Series, Model	Score	Target population	Predict probability
Decision Tree Mo...	0.89	51.59%	87.32%
Clustering Model ...	0.89	51.71%	86.38%
Neural Network M...	0.88	51.00%	86.49%
Random Guess M...		50.00%	
Ideal Model for: D...		58.89%	

Overall Model :



Lift Score:

Mining Legend

Population percentage: 50.00%

Series, Model	Score	Population correct	Predict probability
Decision Tree Mo...	0.87	43.80%	87.32%
Clustering Model ...	0.87	43.90%	86.38%
Neural Network M...	0.86	43.30%	86.49%
Ideal Model		50.00%	

Classification Matrix:

Counts for Decision Tree Model DMX on Accident Severity				
	Predicted	2 (Actual)	3 (Actual)	1 (Actual)
	2	0	0	0
	3	140	849	11
	1	0	0	0
Counts for Clustering Model DMX on Accident Severity				
	Predicted	2 (Actual)	3 (Actual)	1 (Actual)
	2	0	0	0
	3	140	849	11
	1	0	0	0
Counts for Neural Network Model DMX on Accident Severity				
	Predicted	2 (Actual)	3 (Actual)	1 (Actual)
	2	0	0	0
	3	140	849	11
	1	0	0	0

Cross-Validation:

Decision Tree Model DMX				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	333	Classification	True Positive	0.000e+000
2	334	Classification	True Positive	0.000e+000
3	333	Classification	True Positive	0.000e+000
			Average	0.000e+000
			Standard Deviation	0.000e+000
1	333	Classification	False Positive	0.000e+000
2	334	Classification	False Positive	0.000e+000
3	333	Classification	False Positive	0.000e+000
			Average	0.000e+000
			Standard Deviation	0.000e+000
1	333	Classification	True Negative	290
2	334	Classification	True Negative	290
3	333	Classification	True Negative	290
			Average	290
			Standard Deviation	0.000e+000
1	333	Classification	False Negative	43
2	334	Classification	False Negative	44
3	333	Classification	False Negative	43
			Average	43.334
			Standard Deviation	0.4716
1	333	Likelihood	Log Score	-0.4491
2	334	Likelihood	Log Score	-0.448
3	333	Likelihood	Log Score	-0.4397
			Average	-0.4456
			Standard Deviation	0.0042
1	333	Likelihood	Lift	-0.001
2	334	Likelihood	Lift	0.0048
3	333	Likelihood	Lift	0.0084
			Average	0.0041
			Standard Deviation	0.0039
1	333	Likelihood	Root Mean Square Error	0.1529

2	334	Likelihood	Root Mean Square Error	0.1483
3	333	Likelihood	Root Mean Square Error	0.1506
			Average	0.1506
			Standard Deviation	0.0019
Clustering Model DMX				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	333	Classification	True Positive	0.000e+000
2	334	Classification	True Positive	0.000e+000
3	333	Classification	True Positive	0.000e+000
			Average	0.000e+000
			Standard Deviation	0.000e+000
1	333	Classification	False Positive	0.000e+000
2	334	Classification	False Positive	0.000e+000
3	333	Classification	False Positive	0.000e+000
			Average	0.000e+000
			Standard Deviation	0.000e+000
1	333	Classification	True Negative	290
2	334	Classification	True Negative	290
3	333	Classification	True Negative	290
			Average	290
			Standard Deviation	0.000e+000
1	333	Classification	False Negative	43
2	334	Classification	False Negative	44
3	333	Classification	False Negative	43
			Average	43.334
			Standard Deviation	0.4716
1	333	Likelihood	Log Score	-0.4459
2	334	Likelihood	Log Score	-0.4823
3	333	Likelihood	Log Score	-0.484
			Average	-0.4707
			Standard Deviation	0.0175
1	333	Likelihood	Lift	-0.0098
2	334	Likelihood	Lift	-0.0294

3	333	Likelihood	Lift	-0.0359
			Average	-0.025
			Standard Deviation	0.0111
1	333	Likelihood	Root Mean Square Error	0.1566
2	334	Likelihood	Root Mean Square Error	0.1561
3	333	Likelihood	Root Mean Square Error	0.1572
			Average	0.1567
			Standard Deviation	0.0005
Neural Network Model DMX				
Partition Index	**Partition Size**	**Test**	**Measure**	**Value**
1	333	Classification	True Positive	43
2	334	Classification	True Positive	44
3	333	Classification	True Positive	43
			Average	43.334
			Standard Deviation	0.4716
1	333	Classification	False Positive	290
2	334	Classification	False Positive	290
3	333	Classification	False Positive	290
			Average	290
			Standard Deviation	0.000e+000
1	333	Classification	True Negative	0.000e+000
2	334	Classification	True Negative	0.000e+000
3	333	Classification	True Negative	0.000e+000
			Average	0.000e+000
			Standard Deviation	0.000e+000
1	333	Classification	False Negative	0.000e+000
2	334	Classification	False Negative	0.000e+000
3	333	Classification	False Negative	0.000e+000
			Average	0.000e+000
			Standard Deviation	0.000e+000
1	333	Likelihood	Log Score	0.000e+000
2	334	Likelihood	Log Score	0.000e+000
3	333	Likelihood	Log Score	0.000e+000

			Average	0.000e+000
			Standard Deviation	0.000e+000
1	333	Likelihood	Lift	0.000e+000
2	334	Likelihood	Lift	0.000e+000
3	333	Likelihood	Lift	0.000e+000
			Average	0.000e+000
			Standard Deviation	0.000e+000
1	333	Likelihood	Root Mean Square Error	NaN
2	334	Likelihood	Root Mean Square Error	NaN
3	333	Likelihood	Root Mean Square Error	NaN
			Average	NaN
			Standard Deviation	NaN

- Base your data mining findings on the best performing algorithm.

Answer:

We see that the Decision Tree, Clustering Model, Neural Network all performed better. But if we see the individual states for the target variables, Decision Tree is the best-performed algorithm of all three based on the overall Lift Score.

One Page Summary for Decision Tree Model:

Considering the Decision Tree Model being perfect of all 3 models came from the point of lift scores being more for all the values of the target variable Accident_Severity.

Considering the 20% Sample for all target states:
When considering the sample Decision Tree included the Model, Model has 17.5% of the population included in it. Also, I was quite comparable with Clustering Model as both performed well in this case.

Mining Legend

Population percentage: 20.00%

Series, Model	Score	Population cor...	Predi...
Decision Tree Mo...	0.87	17.50%	88.38...
Clustering Model ...	0.87	17.60%	87.22...
Neural Network M...	0.86	16.80%	89.80...
Ideal Model		20.00%	

Classification Matrix:

Counts for Decision Tree Model DMX on Accident Severity:

Predicted	2 (Actual)	3 (Actual)	1 (Actual)
2	0	0	0
3	140	849	11
1	0	0	0

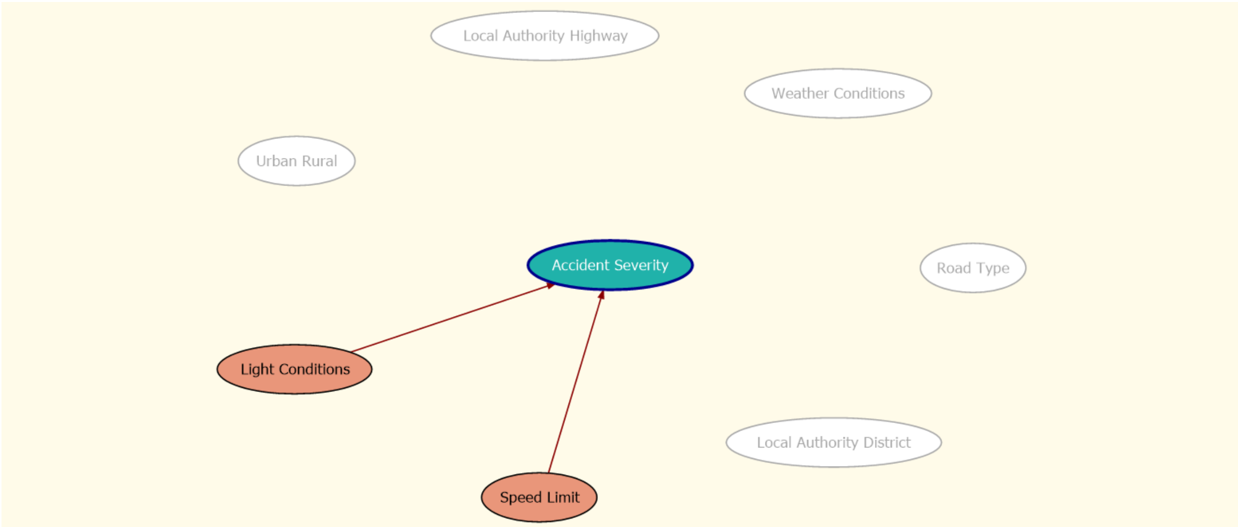
For the Decision Tree, the classification seems to have a more accurate prediction for the target state of 3 as the as there are more data present in the target state of 3(Slight) accident severity. And the model has an 85% classification rate.

Cross-Validation:

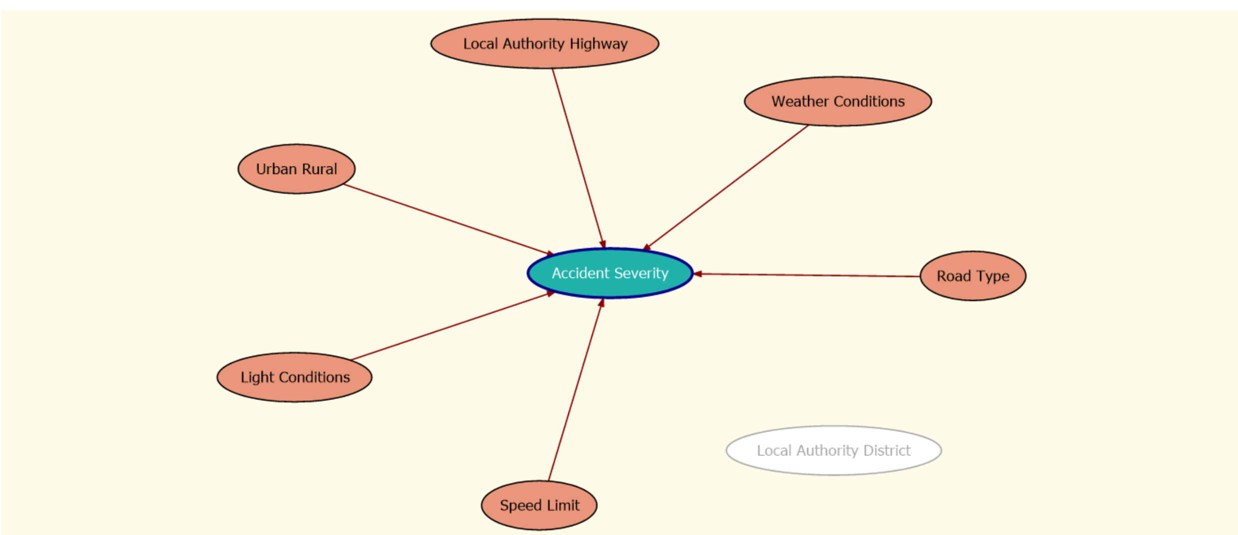
True Negative Cases: 290
False Negative Cases: 43

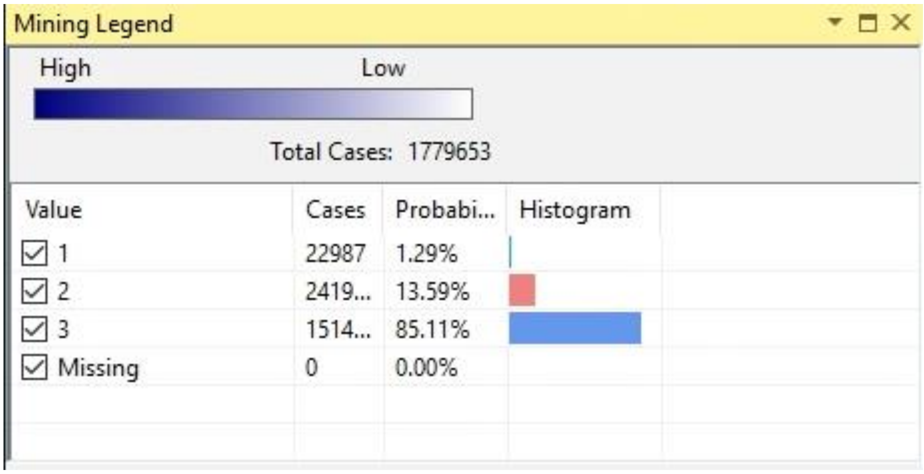
Decision Tree has the lowest false Negative cases compared to all other models and has the highest lift score of 0.61 for a state of 2(Serious),0.89 for a target state of 3, and 0.71 for a target state of 1 for the target variable Accident_Severity.

Dependency Network for Decision Tree:



In this Model, Speed Limit is highly significant compared to all other input variables, and the Local authority District being the least significant in the Model.





The above legend clearly states that the data has most of the target states as 3 and hence the models are more significant for the target state of 3 than for other states. From all these, we can say that the Model is the best suit for prediction of Slight Severity accidents based on all the results.