Missing Data Frequencies

Legend: .=A, B, etc = Missing

time	Frequency	Percent
Non-missing	2480	100.00

Row	Frequency	Percent
Non-missing	2480	100.00

acci_ref	Frequency	Percent
Non-missing	2480	100.00

loc_east_osgr	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

loc_nor_osgr	Frequency	Percent
Non-missing	2480	100.00

longitude	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

latitude	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

police_force	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

acci_severity	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

num_of_vehi	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

num_of_casu	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

date	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

day_of_week	Frequency	Percent
.	1	0.04
Non-missing	2479	99.96

local_auth_distr	Frequency	Percent
Non-missing	2480	100.00

loc_auth_ons_distr	Frequency	Percent
Non-missing	2480	100.00

loc_auth_highw	Frequency	Percent
	1	0.04
Non-missing	2479	99.96

first_road_class	Frequency	Percent
Non-missing	2480	100.00

first_road_num	Frequency	Percent
Non-missing	2480	100.00

road_type	Frequency	Percent
Non-missing	2480	100.00

speed_limit	Frequency	Percent
Non-missing	2480	100.00

junc_detail	Frequency	Percent
Non-missing	2480	100.00

junc_con	Frequency	Percent
Non-missing	2480	100.00

sec_road_class	Frequency	Percent
Non-missing	2480	100.00

sec_road_num	Frequency	Percent
Non-missing	2480	100.00

ped_cross_hum_con	Frequency	Percent
Non-missing	2480	100.00

ped_cross_phy_facil	Frequency	Percent
Non-missing	2480	100.00

light_con	Frequency	Percent
Non-missing	2480	100.00

weath_con	Frequency	Percent
Non-missing	2480	100.00

road_surf_con	Frequency	Percent
Non-missing	2480	100.00

spec_con_site	Frequency	Percent
Non-missing	2480	100.00

carri_haz	Frequency	Percent
Non-missing	2480	100.00

urb_or_rur_area	Frequency	Percent
Non-missing	2480	100.00

did_poli_offi_att	Frequency	Percent
Non-missing	2480	100.00

tru_road_flag	Frequency	Percent
Non-missing	2480	100.00

lsoa_of_acc_loc	Frequency	Percent
Non-missing	2480	100.00

hour_of_day	Frequency	Percent
Non-missing	2480	100.00

time_category	Frequency	Percent
Non-missing	2480	100.00

Missing Data Patterns across Variables

Legend: .=A, B, etc = Missing

time	Row	acci_ref	loc_east_osgr	loc_nor_osgr	longitude	latitude	police_force	acci_severity	num_of_vehi	num_of_casu	date	day_of_week	local_auth_distr	loc_auth_ons_distr	loc_auth_highw	first_road_class	first_road_num	road_type	spee
Non-missing	Non-missing	Non-missing	.	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.	Non-missing	Non-missing	Non-missing	Non-missing	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.	Non-missing	Non-missing	Non-missing	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.	Non-missing	Non-missing	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.	Non-missing	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.	.
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	.

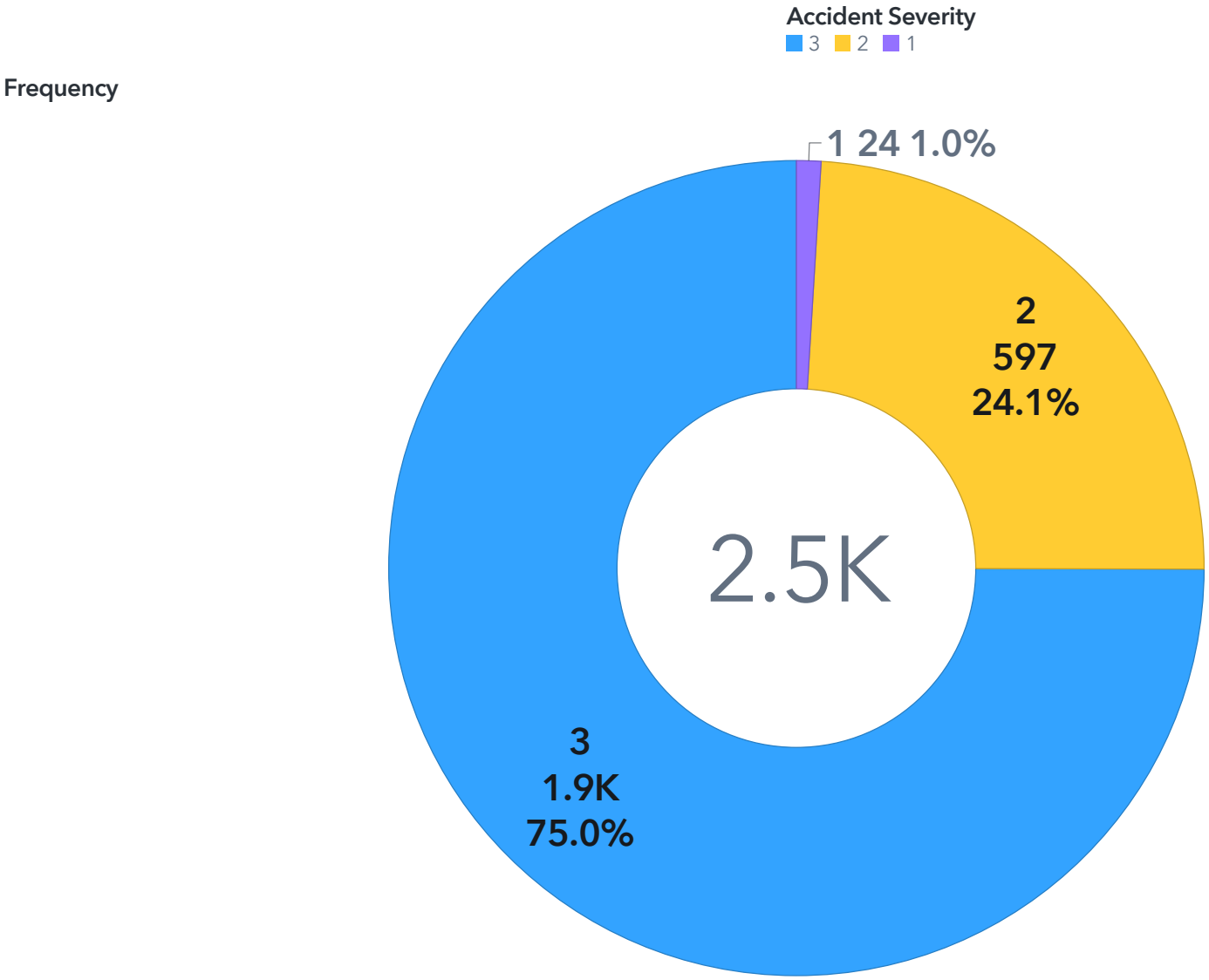
Variable	Mean	Std Dev	Minimum	Maximum	N
time	50215.09	17892.30	60.0000000	86100.00	2480
Row	1240.50	716.0586568	1.0000000	2480.00	2480
acci_ref	451069167	33085.53	451011255	451160434	2480
loc_east_osgr	509570.09	14018.85	482163.00	543673.00	2479
loc_nor_osgr	157127.41	9087.04	132324.00	175208.00	2480
longitude	-0.4296579	0.2006301	-0.8317170	0.0570740	2479
latitude	51.3025884	0.0820440	51.0832120	51.4663730	2479
police_force	45.0000000	0	45.0000000	45.0000000	2479
acci_severity	2.7398144	0.4603663	1.0000000	3.0000000	2479
num_of_vehi	1.9096410	0.7675206	1.0000000	8.0000000	2479
num_of_casu	1.2803550	0.6953533	1.0000000	9.0000000	2479
date	22470.43	101.8460414	22281.00	22645.00	2479
day_of_week	4.0883421	1.9692607	1.0000000	7.0000000	2479
local_auth_distr	-1.0000000	0	-1.0000000	-1.0000000	2480
first_road_class	4.0766129	1.6026707	1.0000000	6.0000000	2480
first_road_num	355.2100806	785.1844174	0	3411.00	2480
road_type	5.1290323	1.6597056	1.0000000	9.0000000	2480
speed_limit	38.8548387	13.8880986	20.0000000	70.0000000	2480
junc_detail	2.0758065	2.9367233	0	9.0000000	2480
junc_con	1.1358871	2.3760271	-1.0000000	4.0000000	2480
sec_road_class	2.3786290	2.7298322	0	6.0000000	2480
sec_road_num	84.4302419	423.7698196	-1.0000000	3411.00	2480
ped_cross_hum_con	0.0237903	0.2102615	0	2.0000000	2480
ped_cross_phy_facil	0.6411290	1.8070059	0	8.0000000	2480
light_con	1.9665323	1.7018868	1.0000000	7.0000000	2480
weath_con	1.5403226	1.6205079	1.0000000	9.0000000	2480
road_surf_con	1.3133065	0.5700473	-1.0000000	5.0000000	2480
spec_con_site	0.1326613	0.7782217	-1.0000000	7.0000000	2480
carri_haz	0.0991935	0.6234617	-1.0000000	7.0000000	2480
urb_or_rur_area	1.4116935	0.4922394	1.0000000	2.0000000	2480
did_poli_offi_att	1.3834677	0.7281363	1.0000000	12.0000000	2480
tru_road_flag	1.8693548	0.3370798	1.0000000	2.0000000	2480
hour_of_day	13.5008065	4.9812071	0	23.0000000	2480

Surrey_VariableNames_Changed

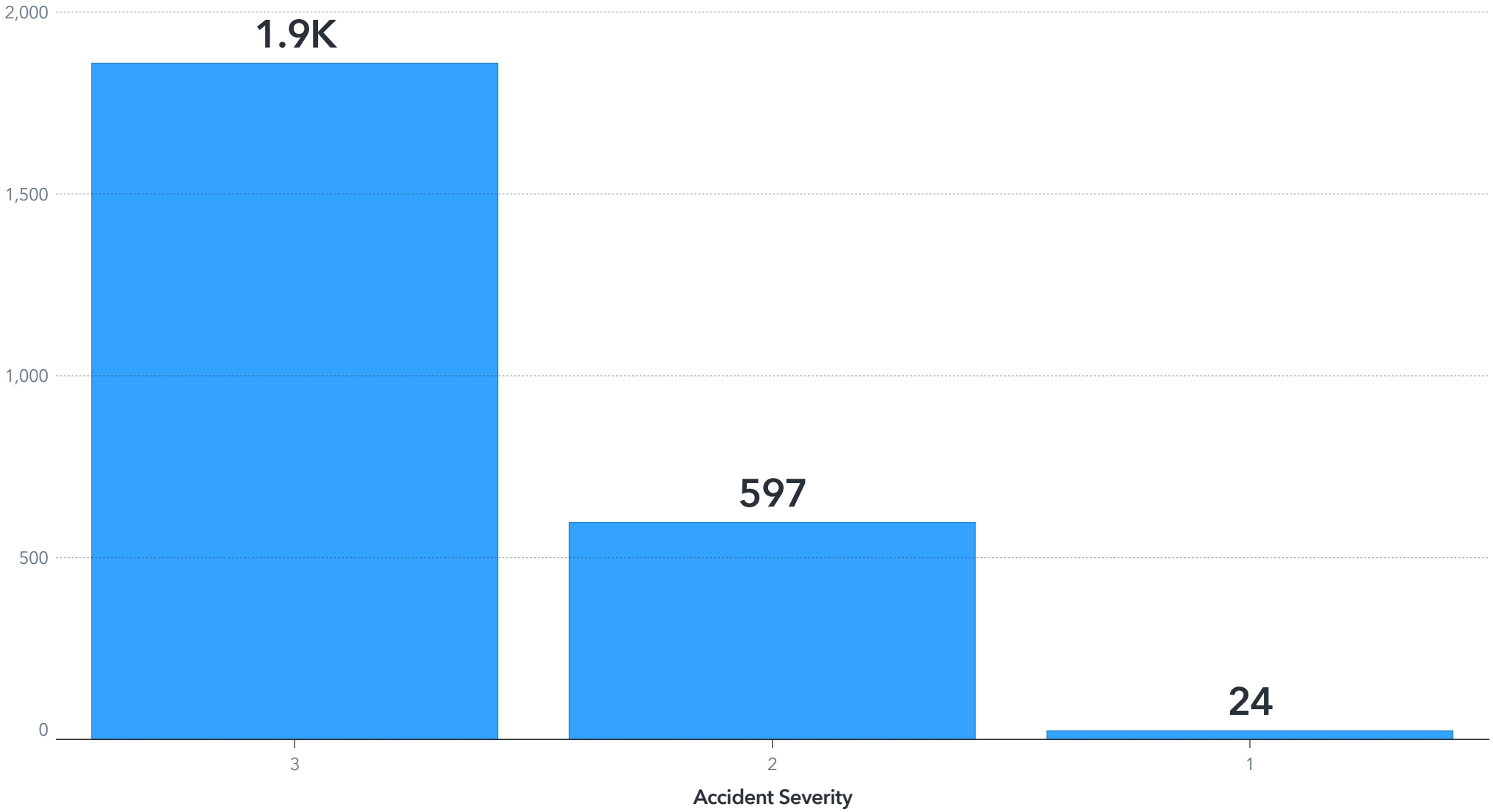
Creation Date: Thursday, 9 January 2025, 21:30:32

Author: ta01468@surrey.ac.uk

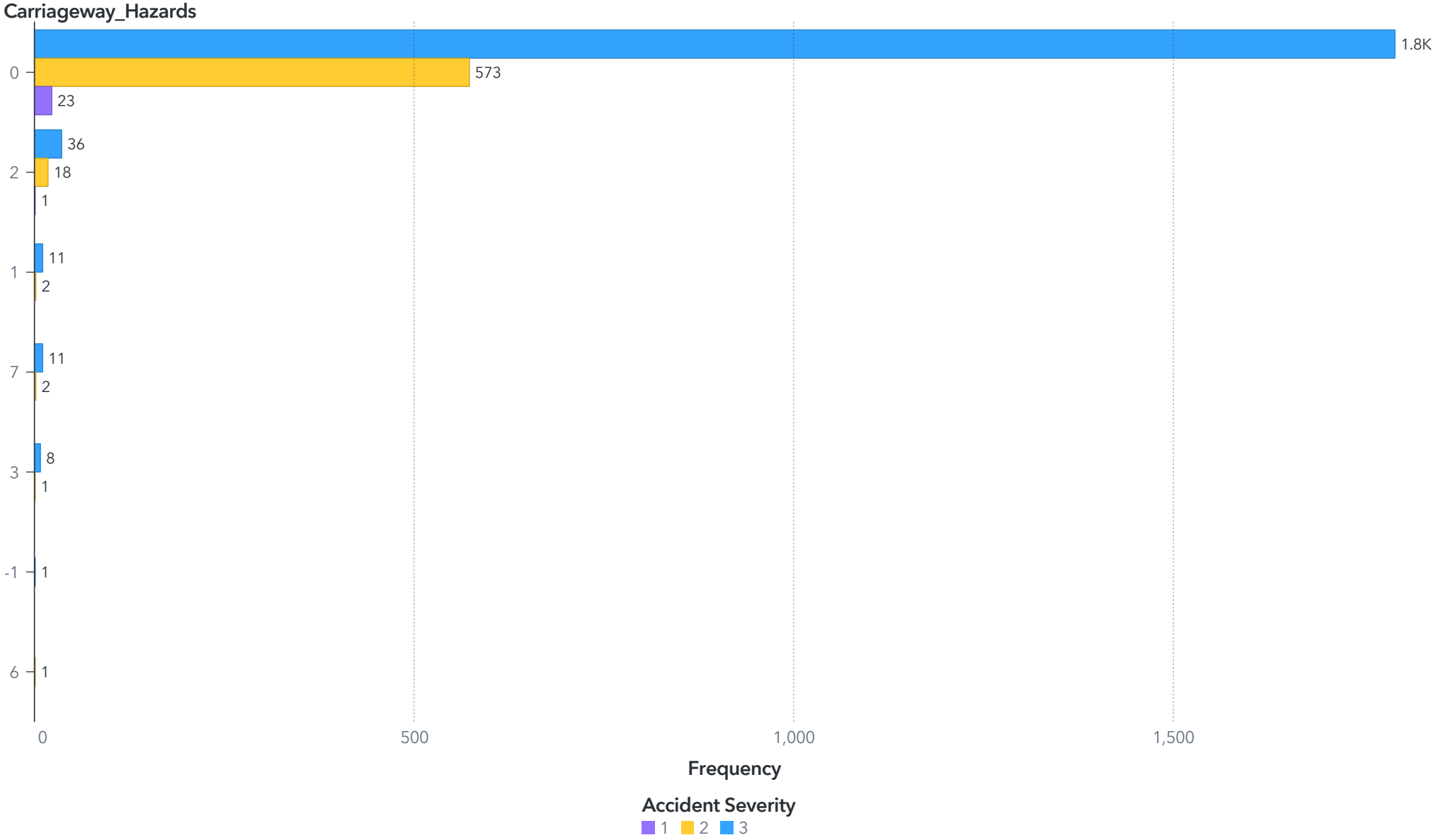
Frequency of Accident Severity

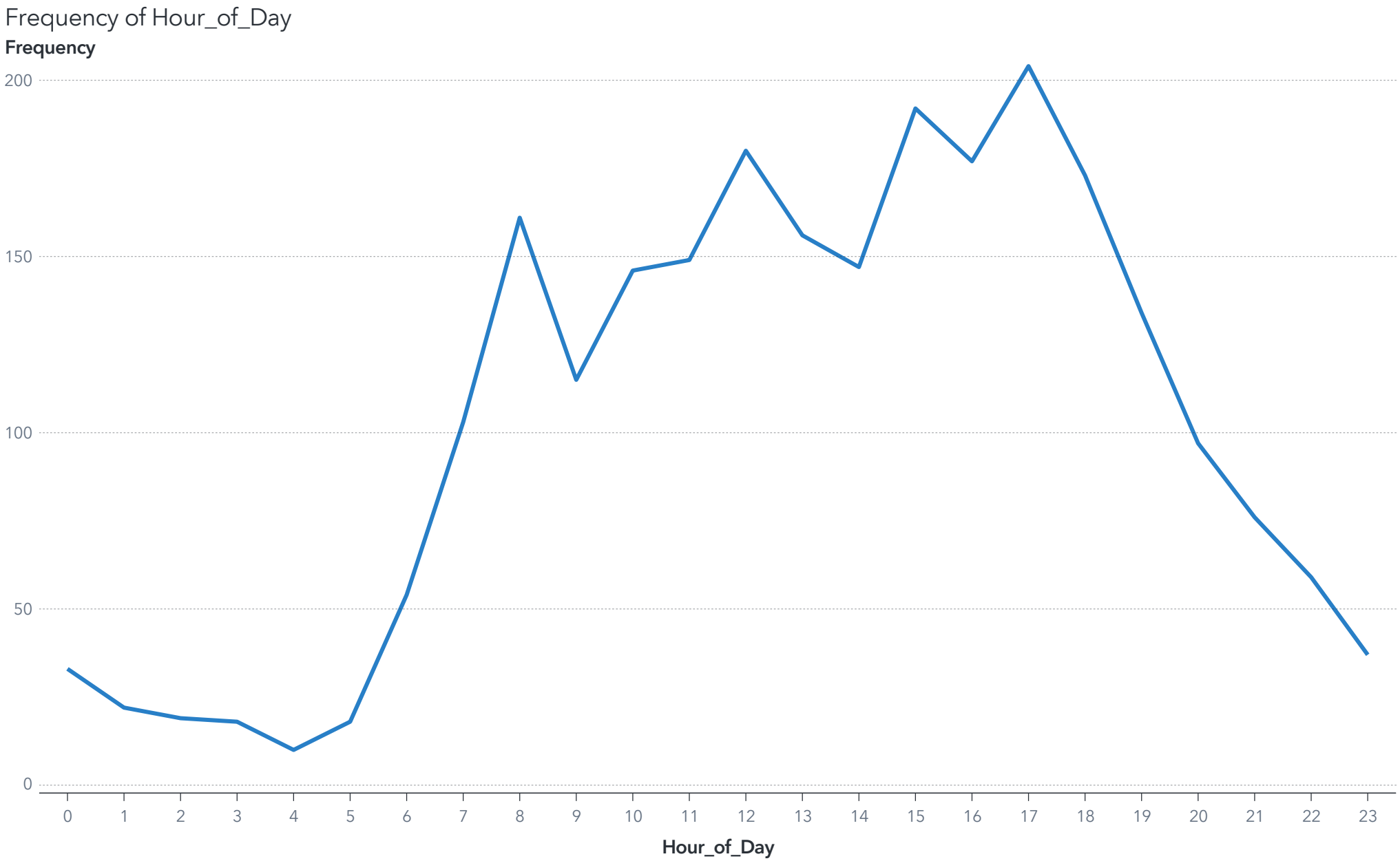


Frequency of Accident Severity

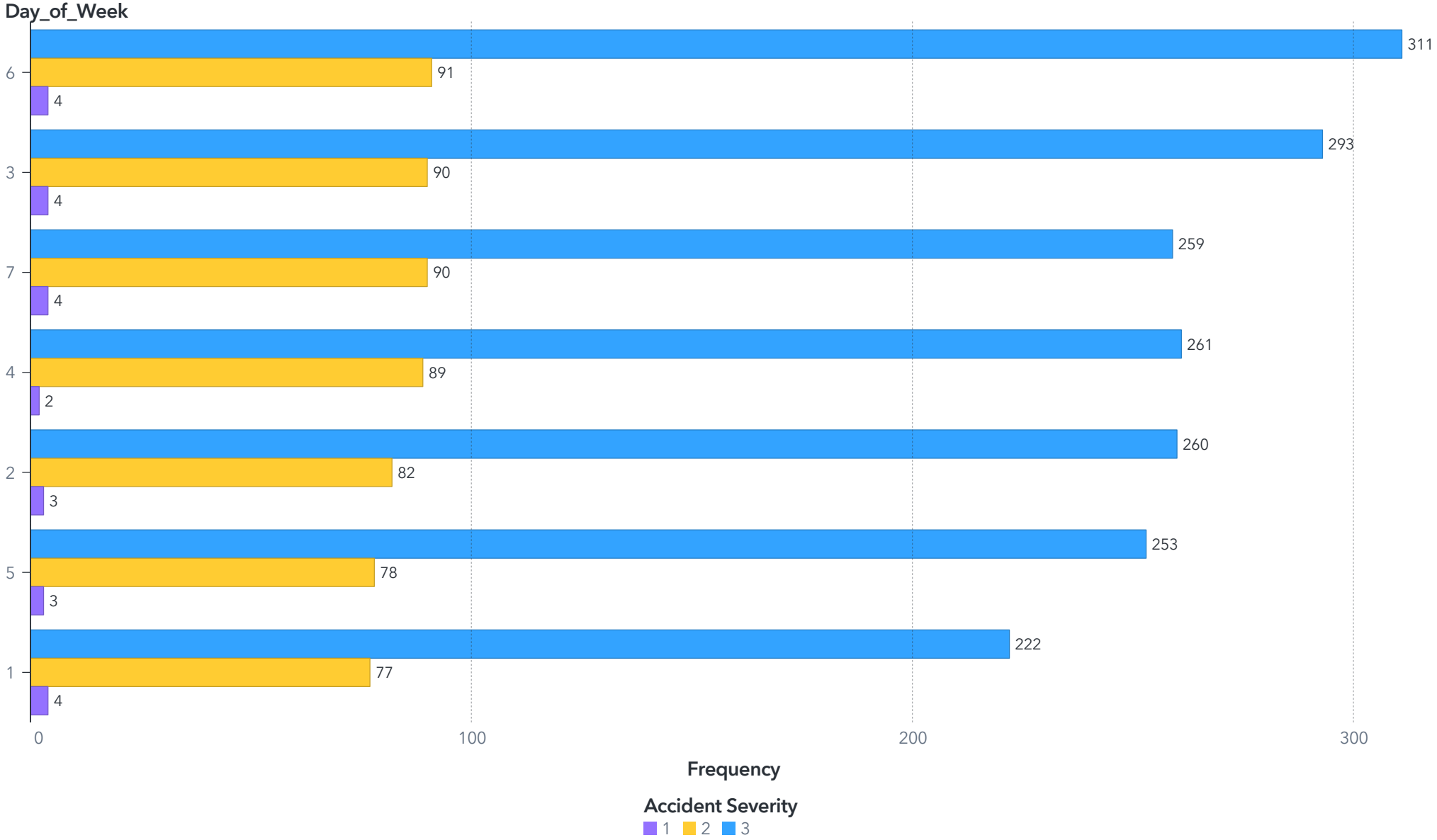


Frequency of Carriageway_Hazards grouped by Accident Severity

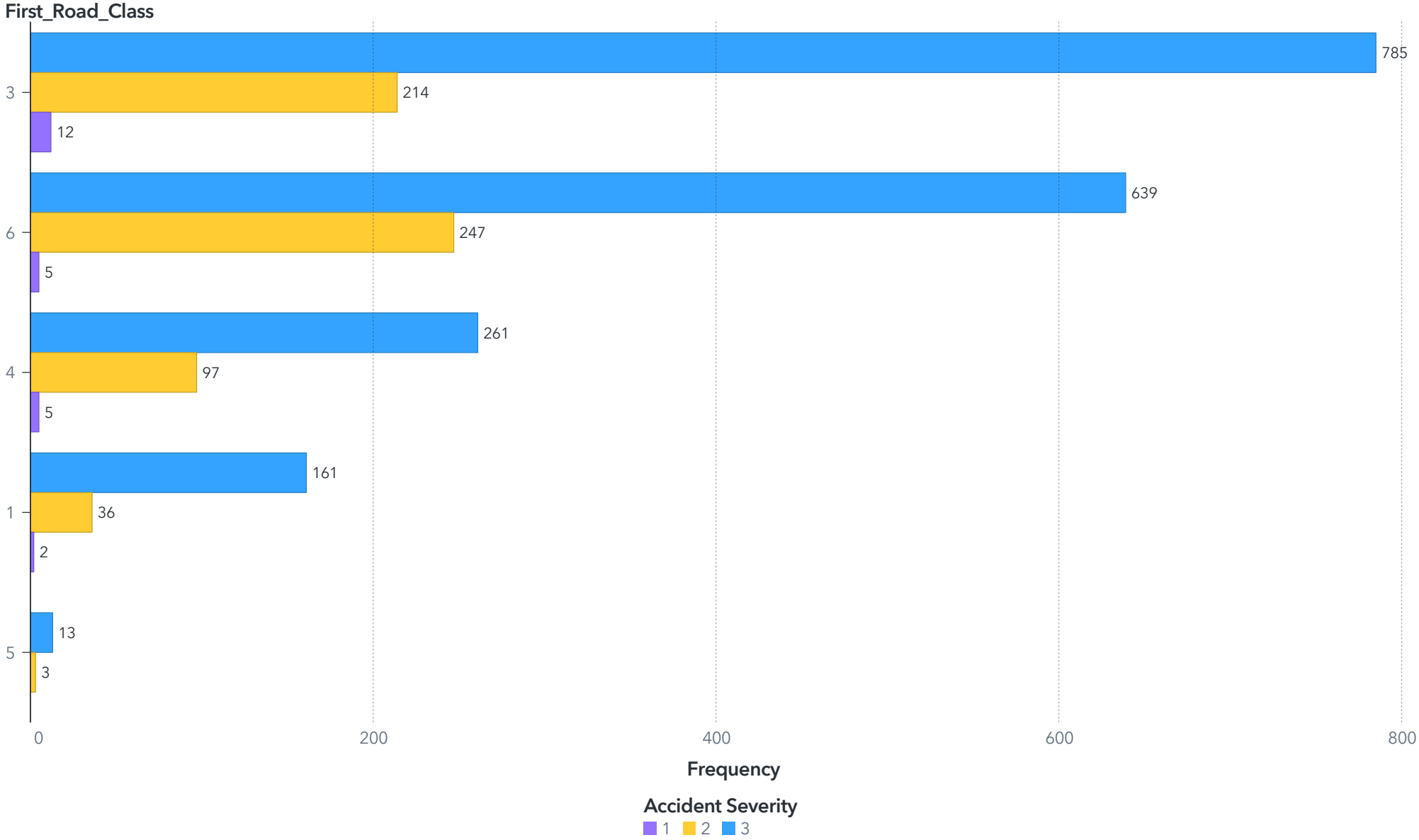



Frequency of Hour_of_Day
Frequency
200
150
100
50
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour_of_Day

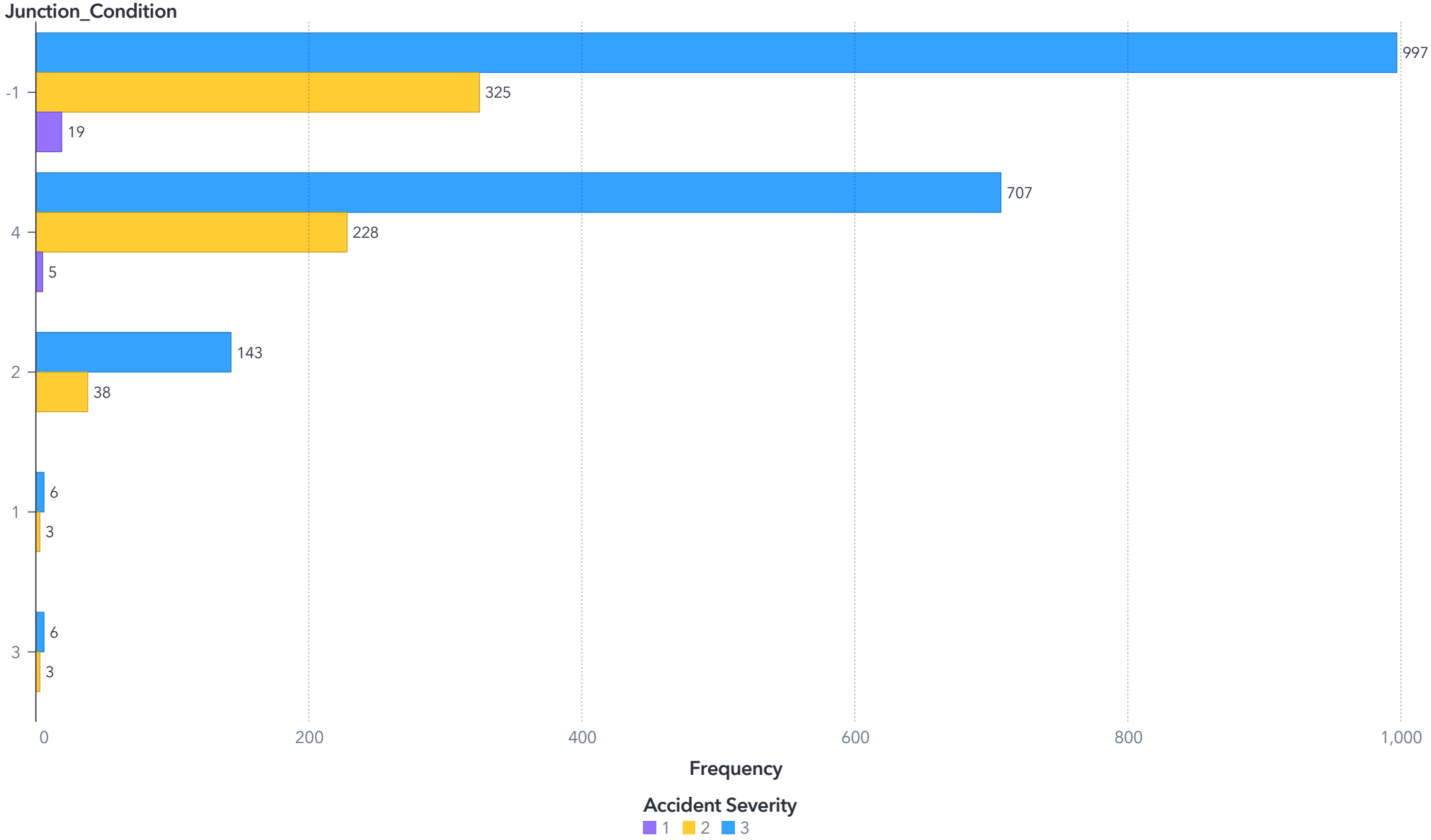
Frequency of Day_of_Week grouped by Accident Severity



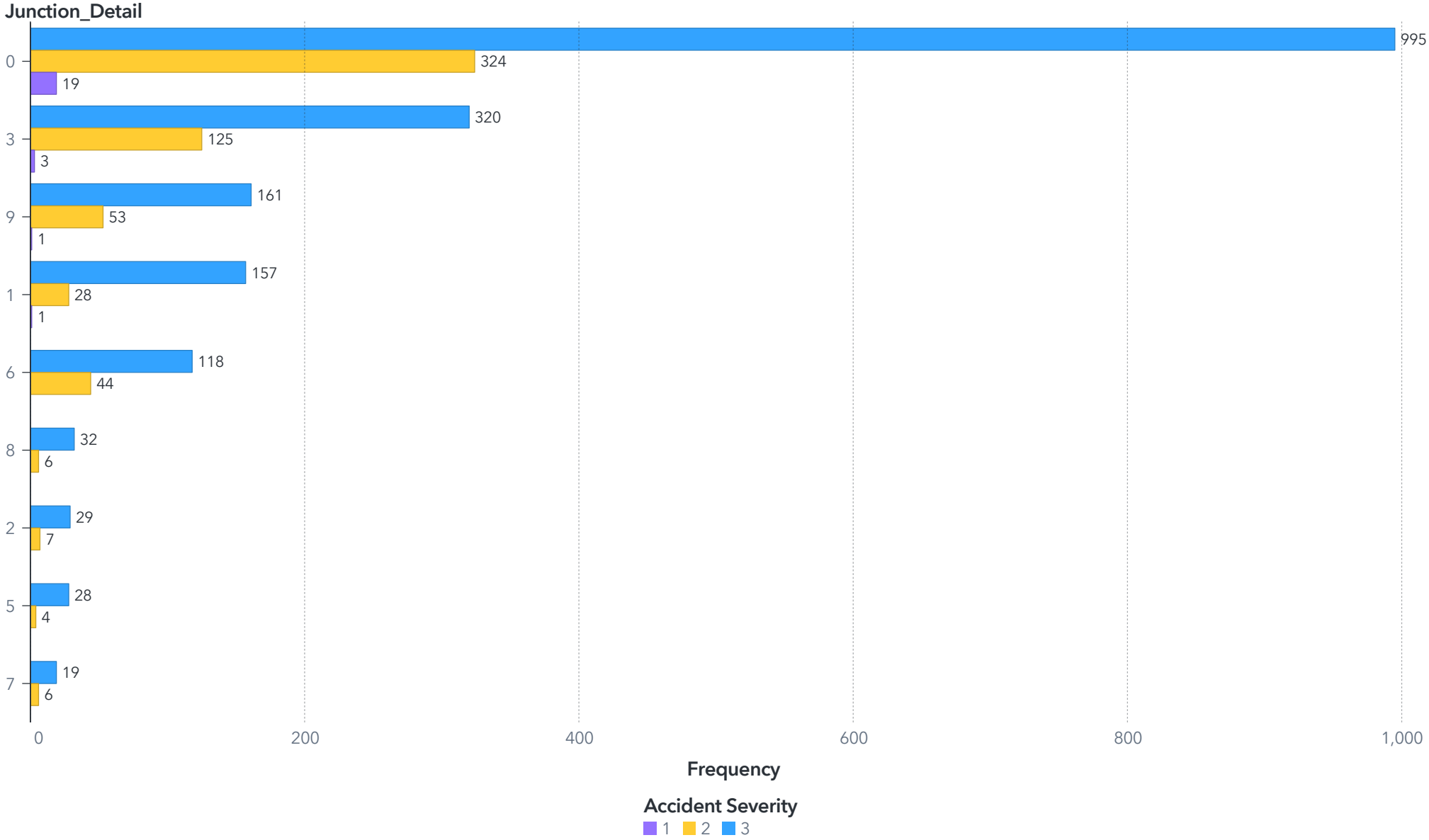
Frequency of First_Road_Class grouped by Accident Severity



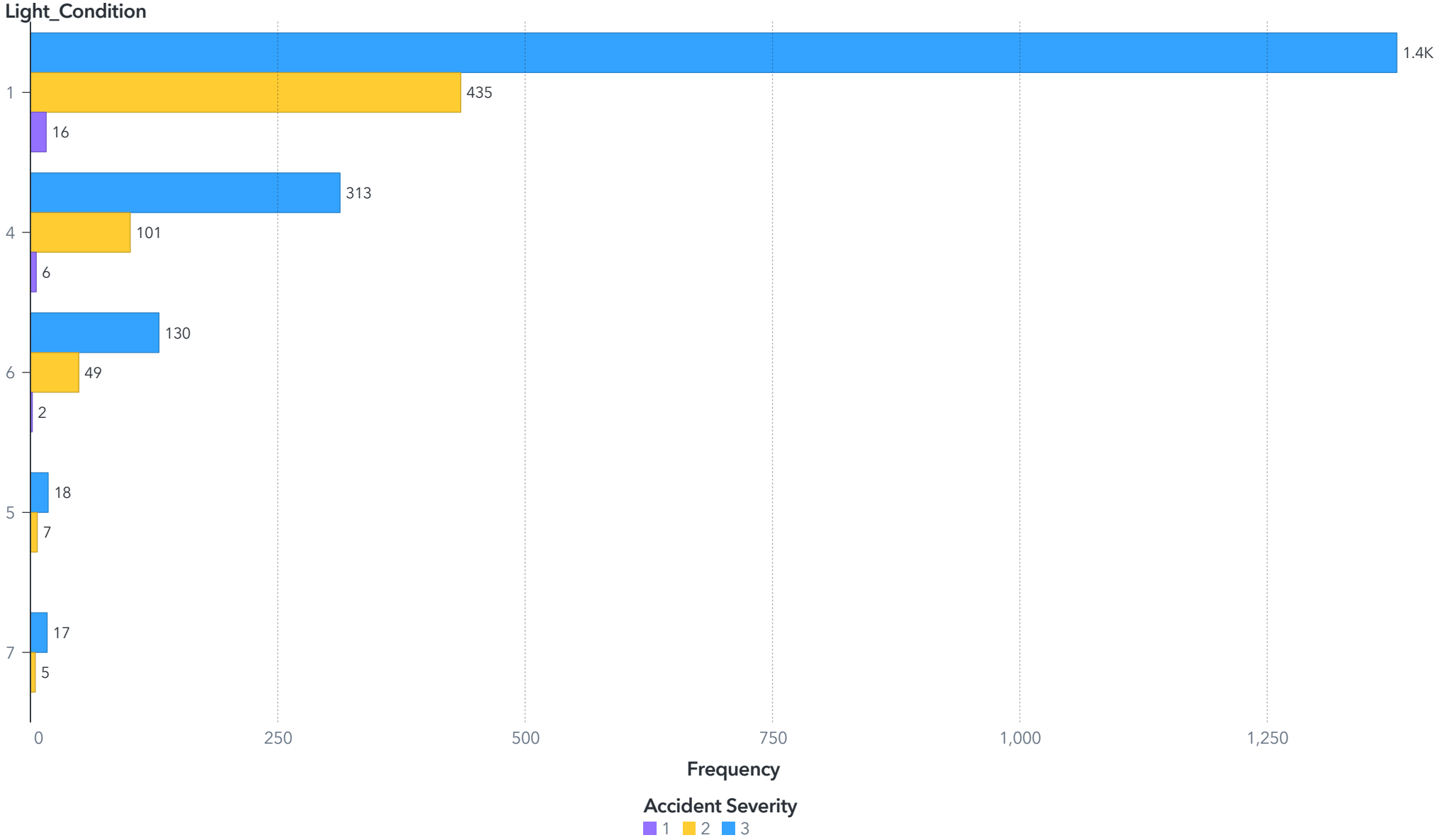
Frequency of Junction_Condition grouped by Accident Severity



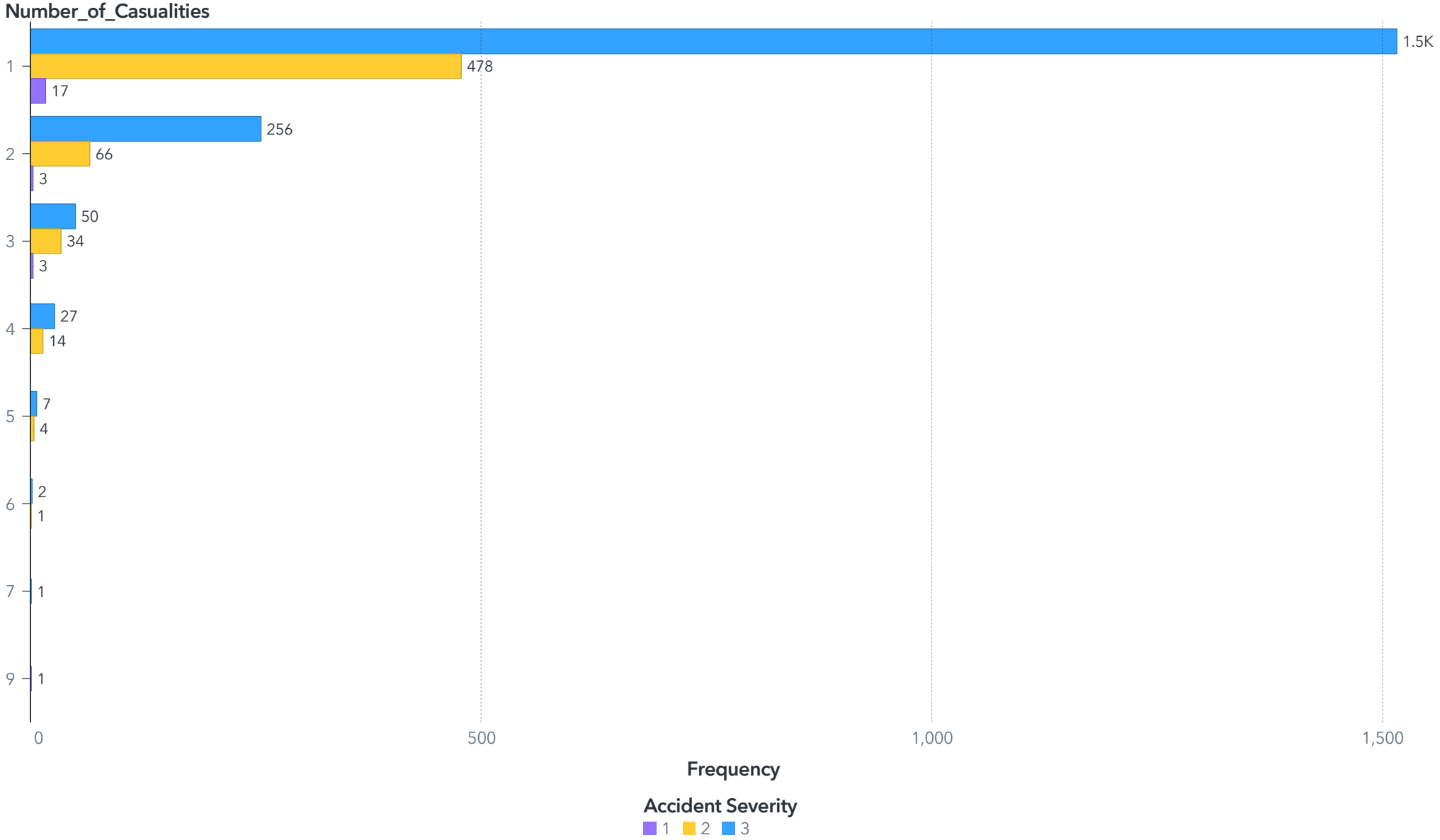
Frequency of Junction_Detail grouped by Accident Severity



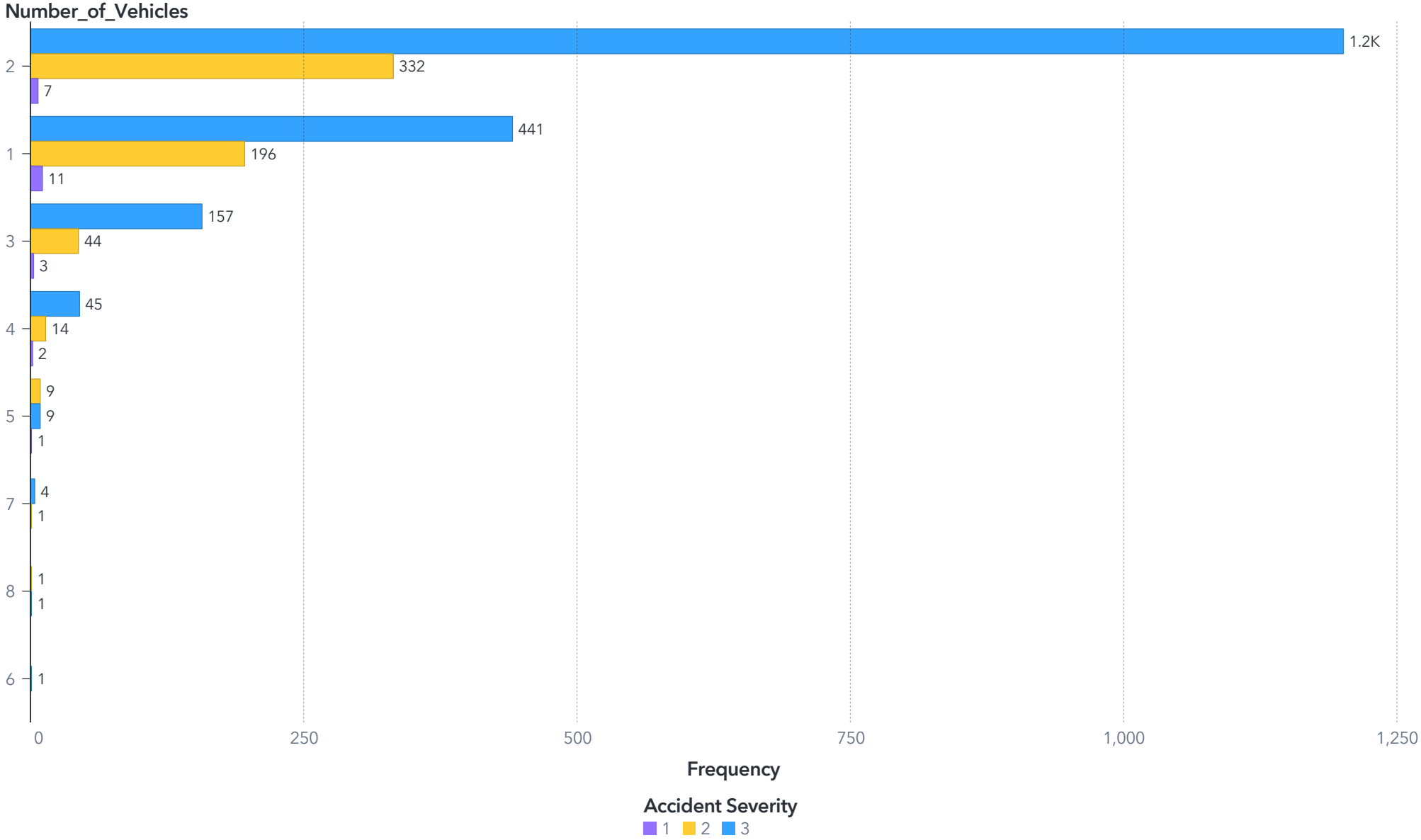
Frequency of Light_Condition grouped by Accident Severity



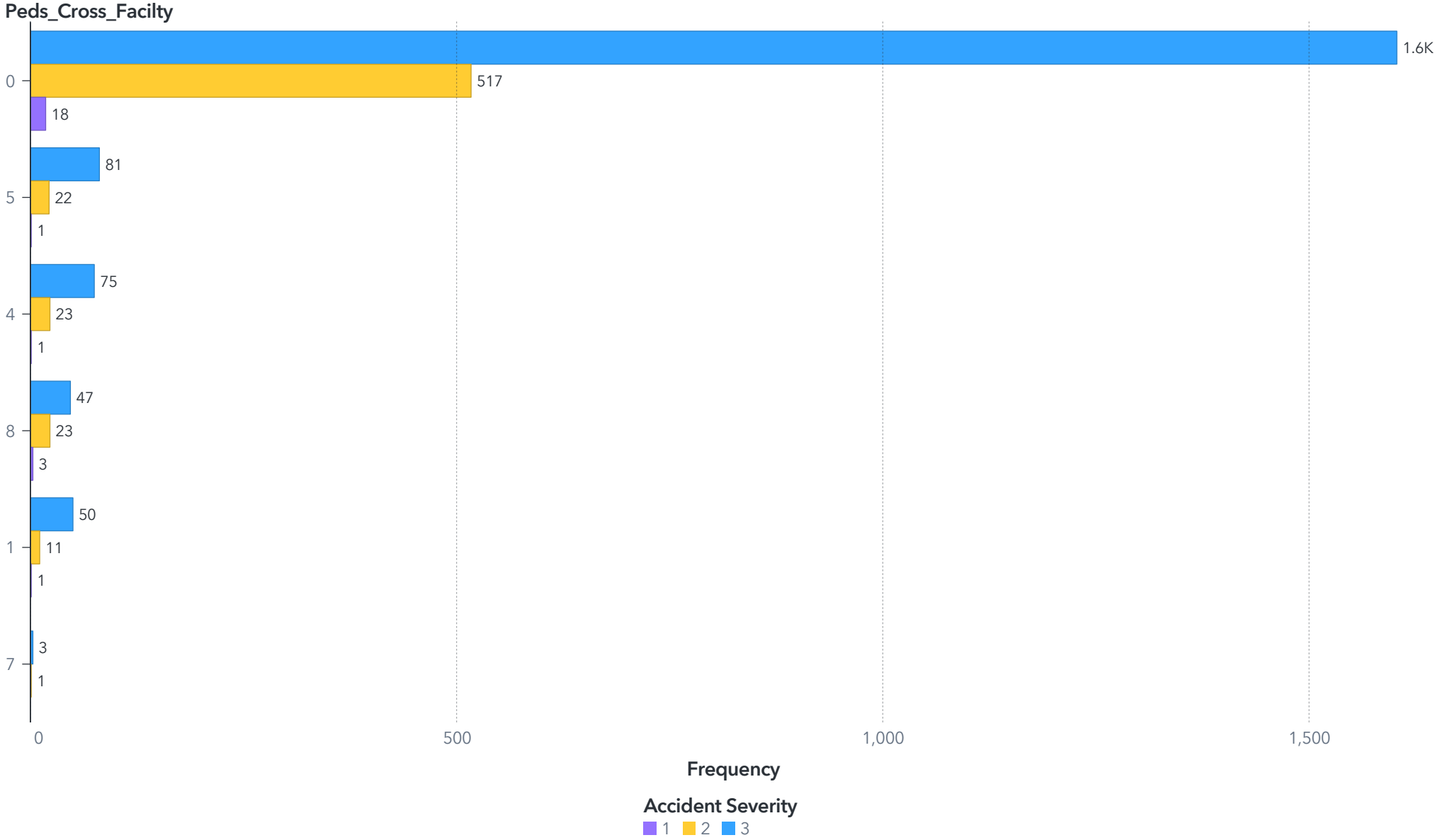
Frequency of Number_of_Casualities grouped by Accident Severity



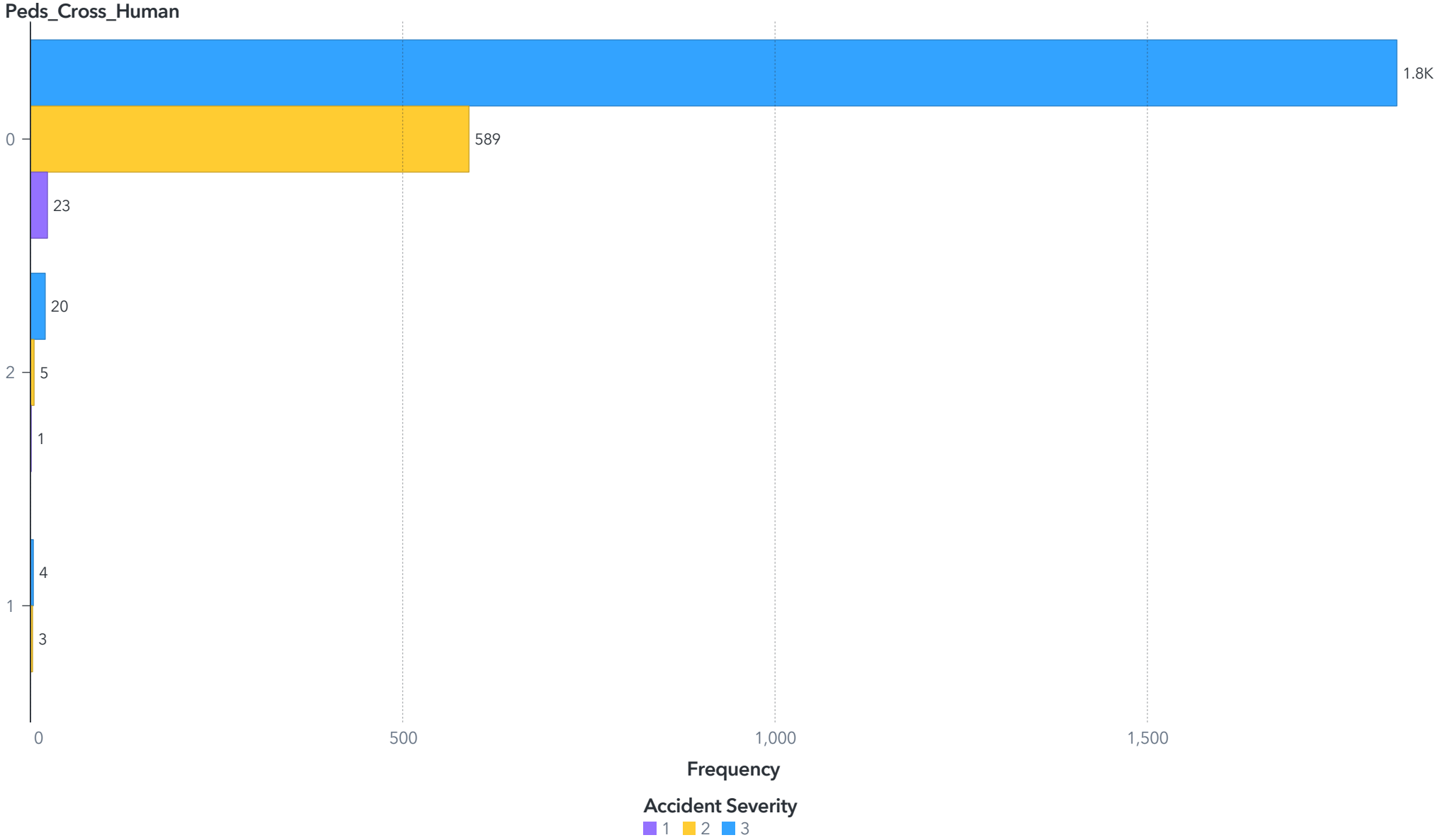
Frequency of Number_of_Vehicles grouped by Accident Severity



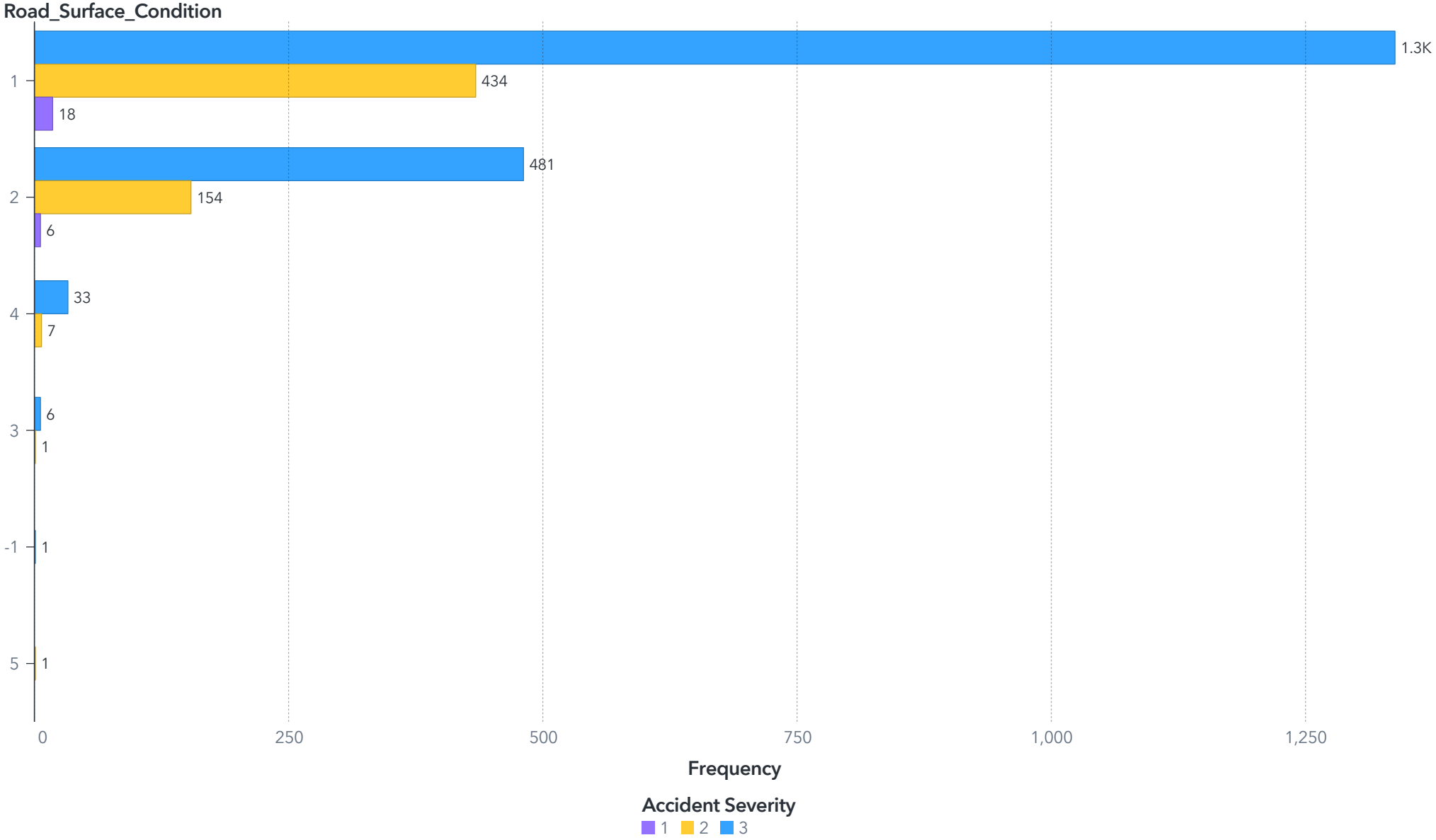
Frequency of Peds_Cross_Facilty grouped by Accident Severity



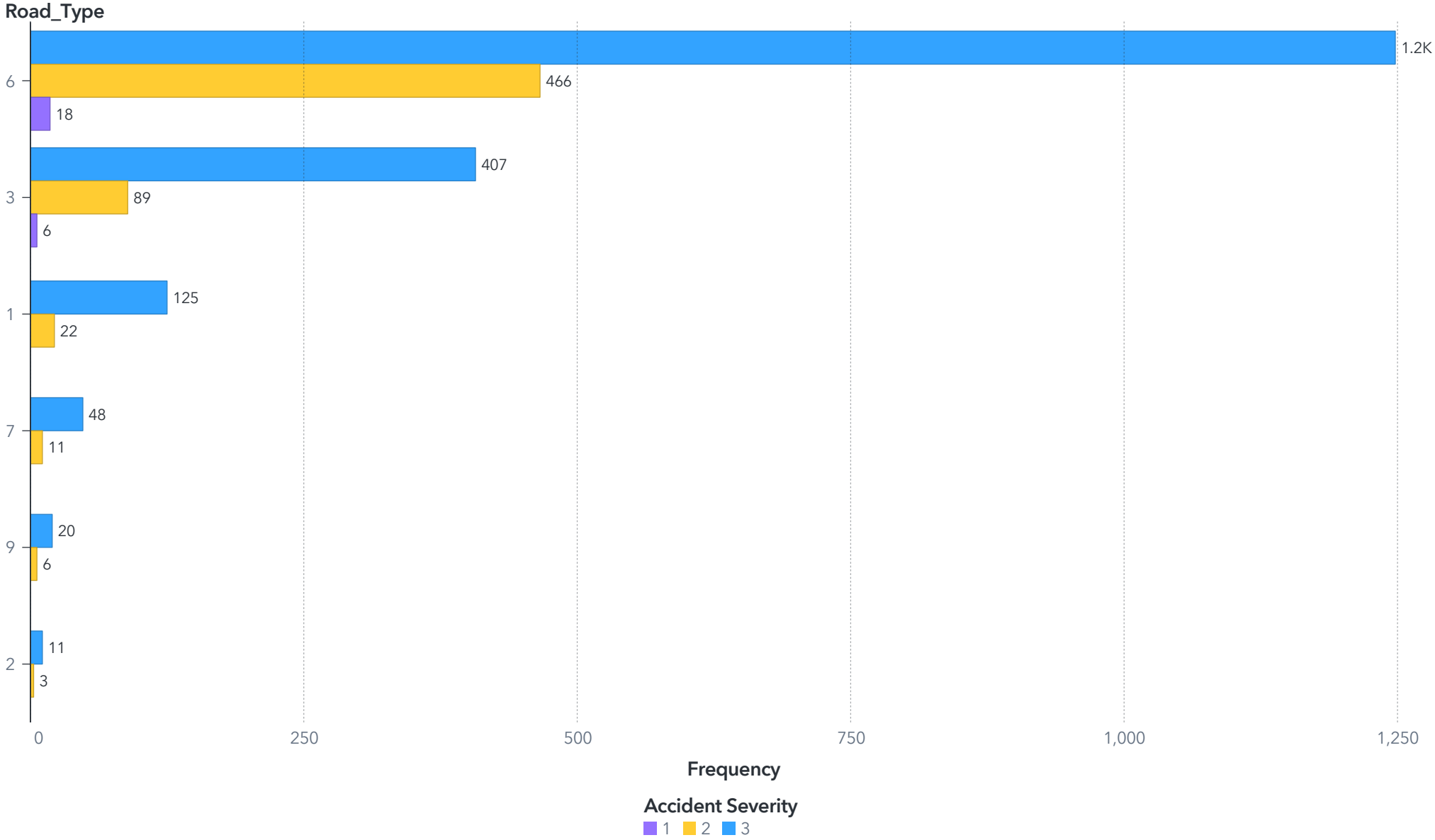
Frequency of Peds_Cross_Human grouped by Accident Severity



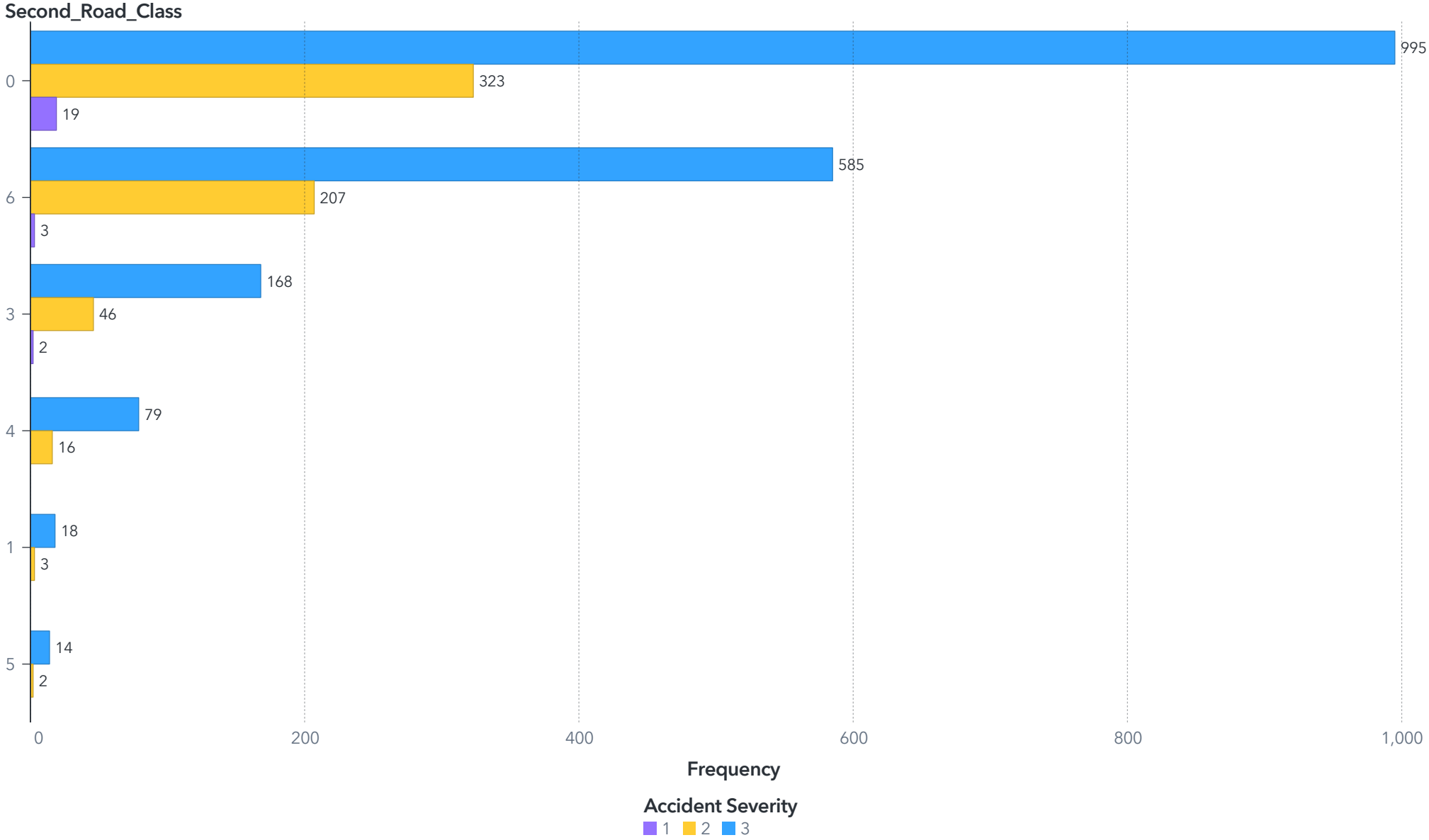
Frequency of Road_Surface_Condition grouped by Accident Severity



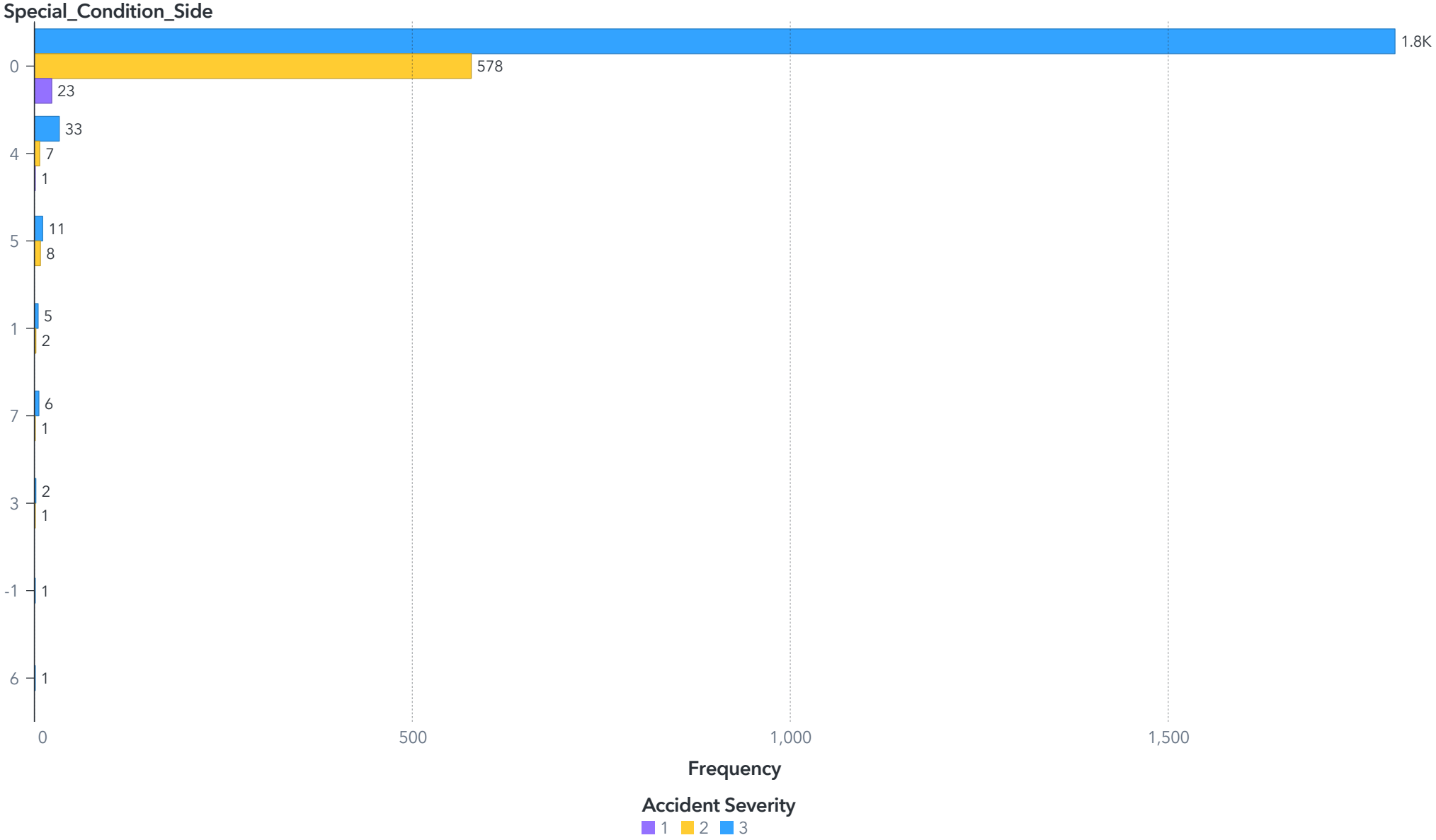
Frequency of Road_Type grouped by Accident Severity



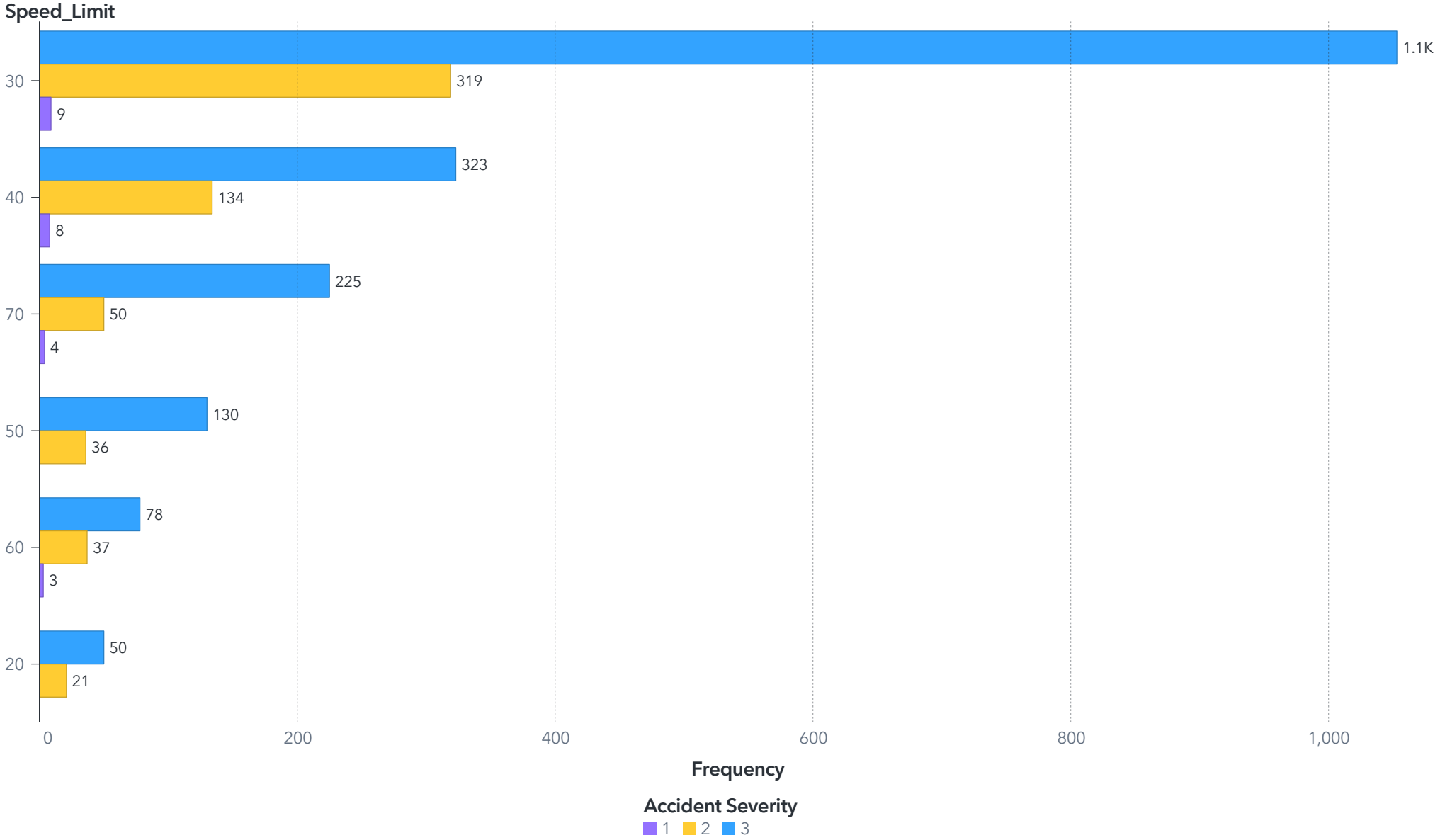
Frequency of Second_Road_Class grouped by Accident Severity



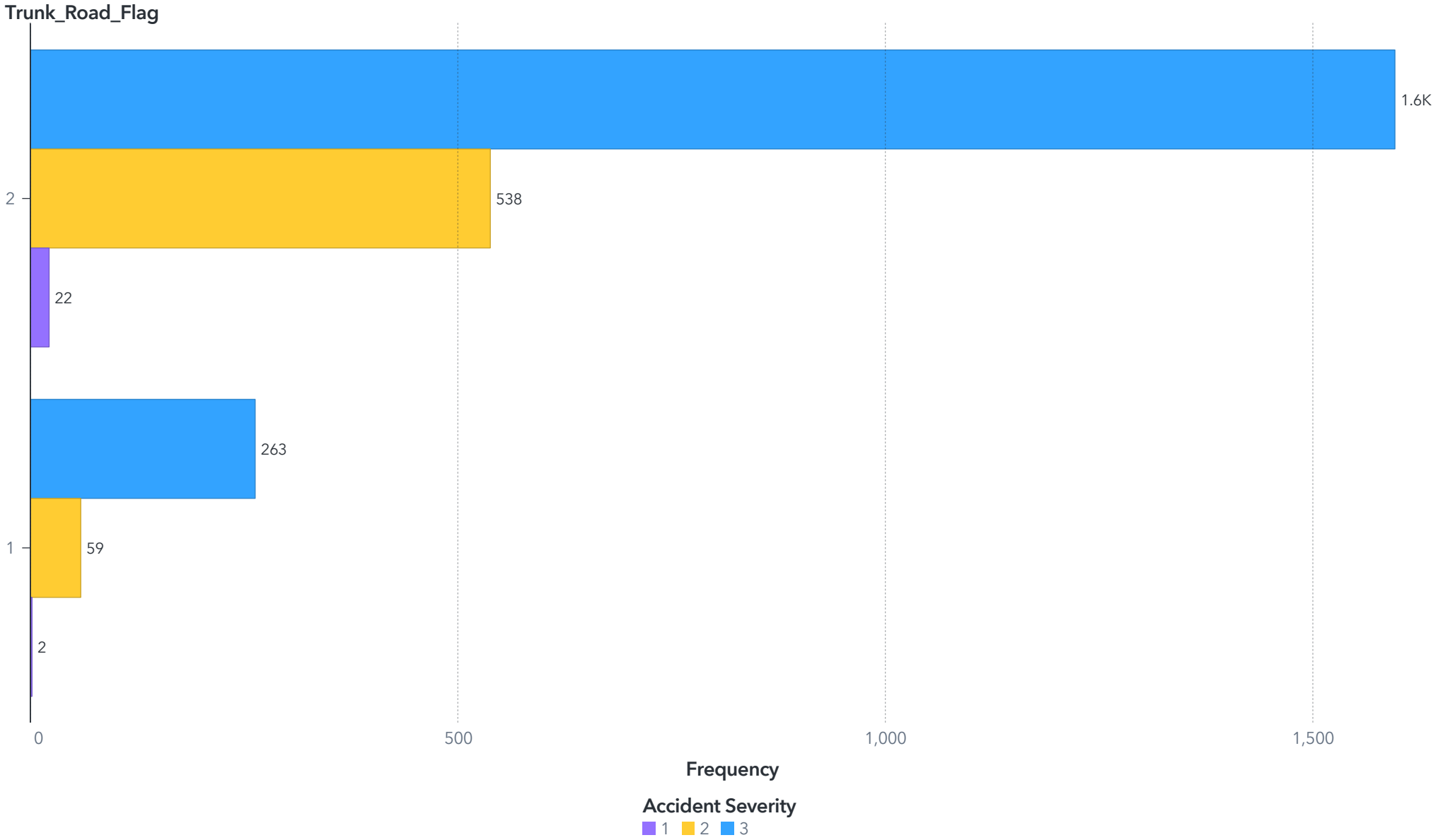
Frequency of Special_Condition_Side grouped by Accident Severity

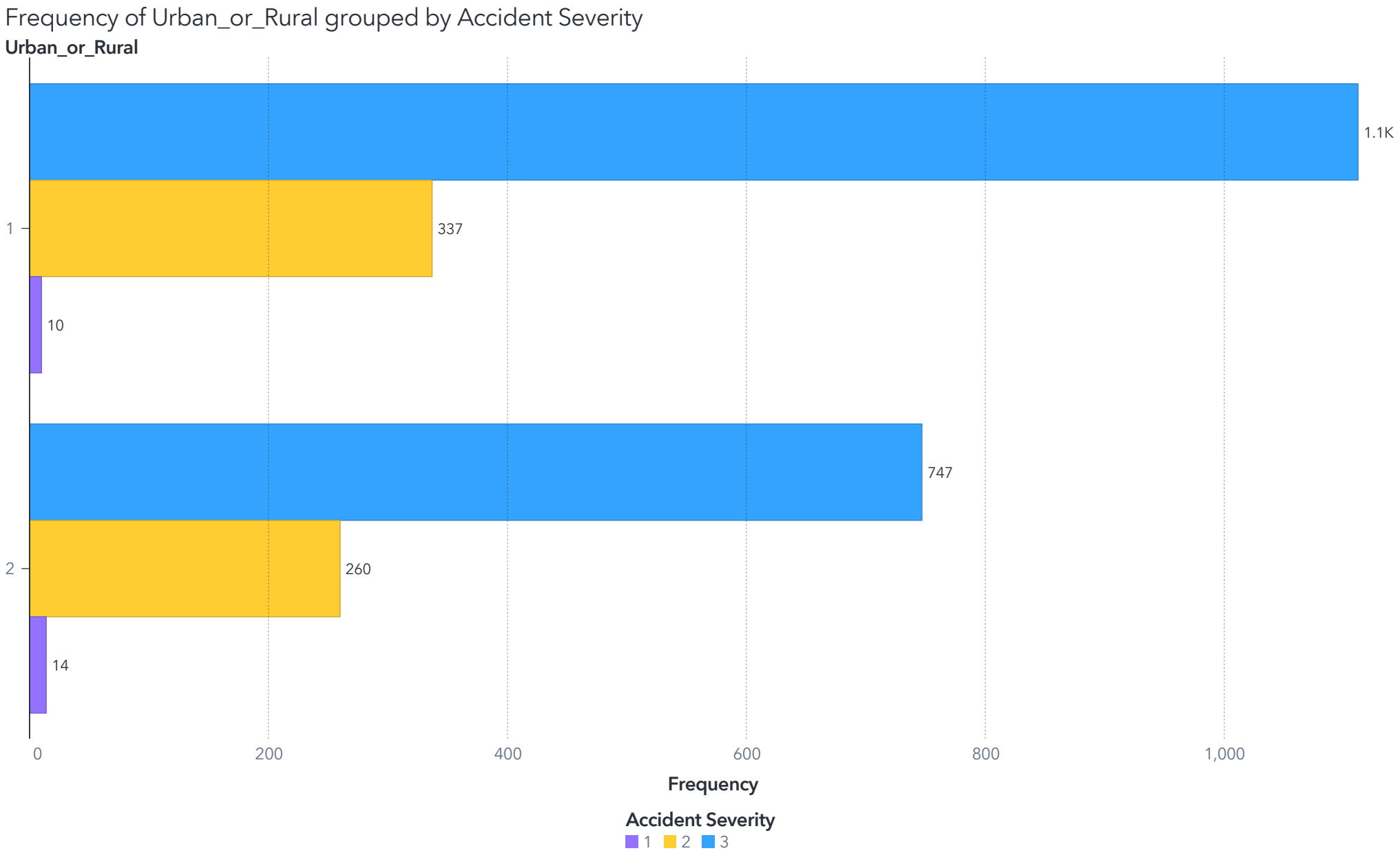


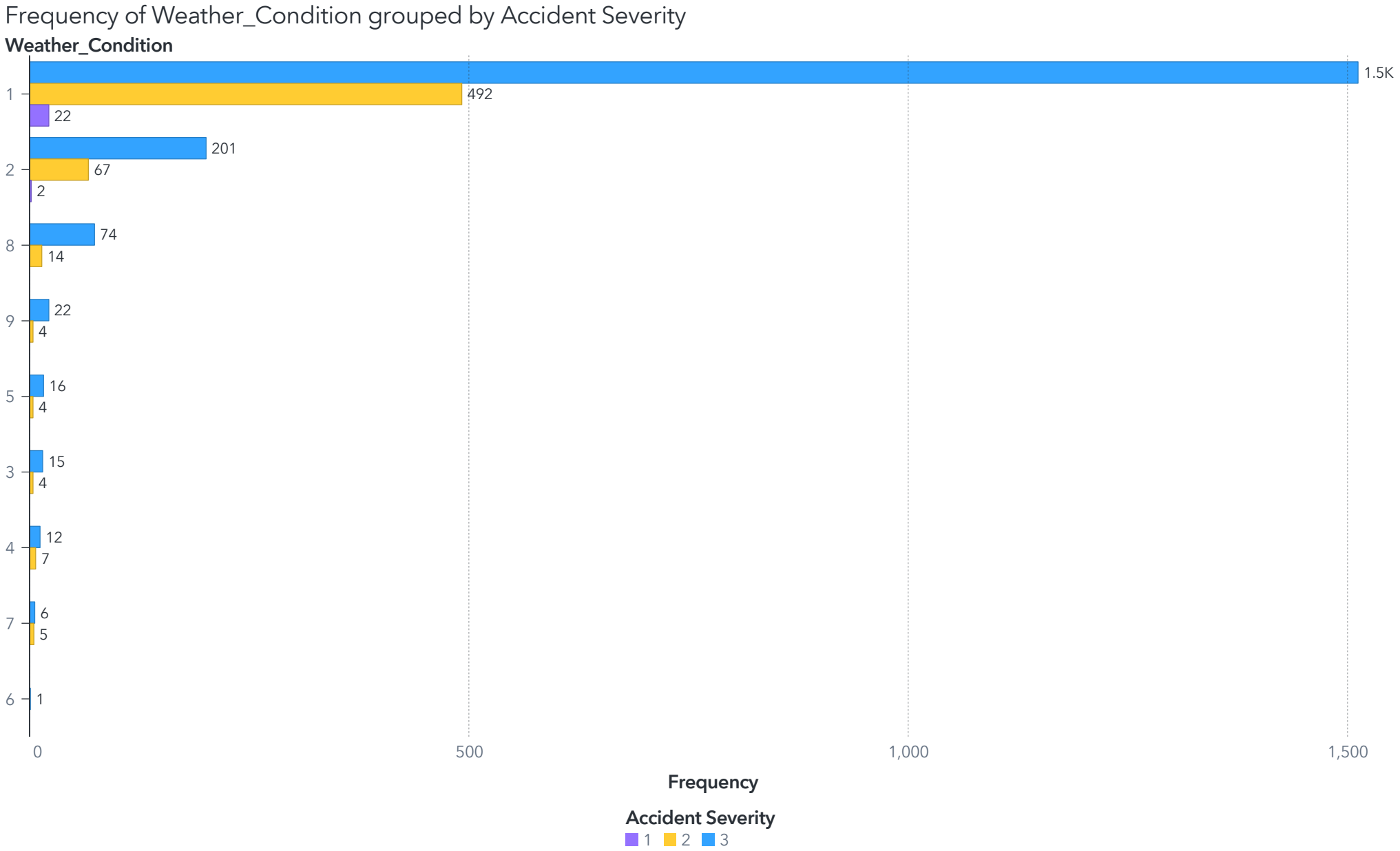
Frequency of Speed_Limit grouped by Accident Severity



Frequency of Trunk_Road_Flag grouped by Accident Severity




Frequency of Urban_or_Rural grouped by Accident Severity
Urban_or_Rural
1
2
1.1K
337
10
747
260
14
0
200
400
600
800
1,000
Frequency
Accident Severity
1
2
3

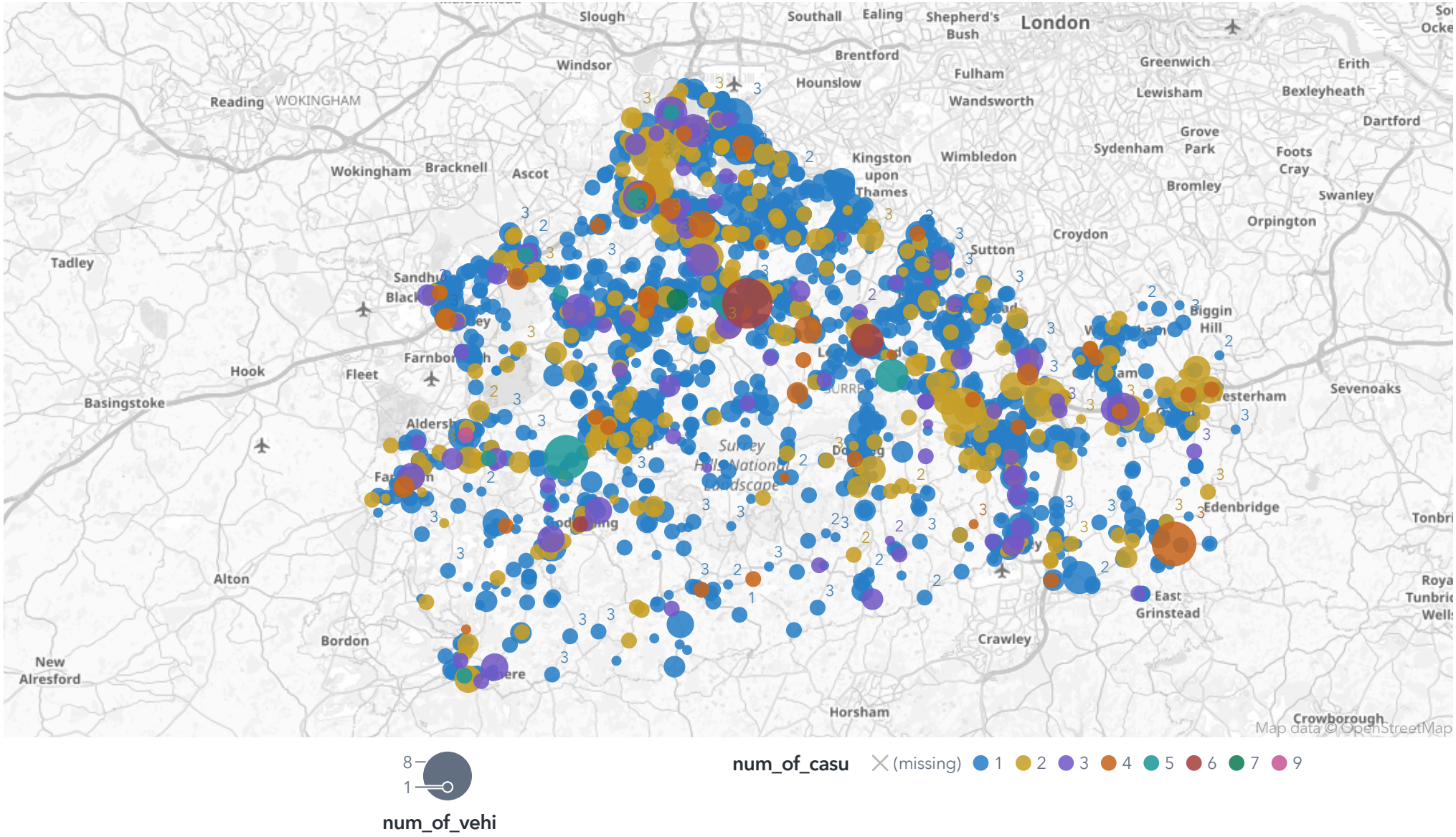

Frequency of Weather_Condition grouped by Accident Severity
Weather_Condition
1
2
8
9
5
3
4
7
6
1.5K
492
22
201
67
2
74
14
22
4
16
4
15
4
12
7
6
5
1
0
500
1,000
1,500
Frequency
Accident Severity
1
2
3

Report 1

Creation Date: Thursday, 9 January 2025, 21:57:01

Author: ta01468@surrey.ac.uk

num_of_casu by Geographic Item 1 sized by num_of_vehi



A1.1

Appendix

A1.1 num_of_casu by Geographic Item 1 sized by num_of_vehi

Warnings: Some features may not be displayed on the map because of missing location information in the data.

Performance Information	
Execution Mode	Single-Machine
Number of Threads	4

Data Access Information			
Data	**Engine**	**Role**	**Path**
WORK.IMPORT	V9	Input	On Client
WORK.HPIMPUTE_SCORES0001	V9	Output	On Client

Imputation Results

Variable	Imputation Indicator	Imputed Variable	N Missing	Type of Imputation	Imputation Value (Seed)
time	**M_time**	**IM_time**	0	Pseudo Median	51660
Row	**M_Row**	**IM_Row**	0	Pseudo Median	1240
acci_ref	**M_acci_ref**	**IM_acci_ref**	0	Pseudo Median	451068264
loc_east_osgr	**M_loc_east_osgr**	**IM_loc_east_osgr**	1	Pseudo Median	507080
loc_nor_osgr	**M_loc_nor_osgr**	**IM_loc_nor_osgr**	0	Pseudo Median	158193
longitude	**M_longitude**	**IM_longitude**	1	Pseudo Median	-0.46401
latitude	**M_latitude**	**IM_latitude**	1	Pseudo Median	51.31257
police_force	**M_police_force**	**IM_police_force**	1	Pseudo Median	45.00000
acci_severity	**M_acci_severity**	**IM_acci_severity**	1	Pseudo Median	3.00000
num_of_vehi	**M_num_of_vehi**	**IM_num_of_vehi**	1	Pseudo Median	2.00000
num_of_casu	**M_num_of_casu**	**IM_num_of_casu**	1	Pseudo Median	1.00000
date	**M_date**	**IM_date**	1	Pseudo Median	22474
day_of_week	**M_day_of_week**	**IM_day_of_week**	1	Pseudo Median	4.00000
local_auth_distr	**M_local_auth_distr**	**IM_local_auth_distr**	0	Pseudo Median	-1.00000
first_road_class	**M_first_road_class**	**IM_first_road_class**	0	Pseudo Median	4.00000
first_road_num	**M_first_road_num**	**IM_first_road_num**	0	Pseudo Median	24.00000
road_type	**M_road_type**	**IM_road_type**	0	Pseudo Median	6.00000
speed_limit	**M_speed_limit**	**IM_speed_limit**	0	Pseudo Median	30.00000
junc_detail	**M_junc_detail**	**IM_junc_detail**	0	Pseudo Median	0
junc_con	**M_junc_con**	**IM_junc_con**	0	Pseudo Median	-1.00000
sec_road_class	**M_sec_road_class**	**IM_sec_road_class**	0	Pseudo Median	0
sec_road_num	**M_sec_road_num**	**IM_sec_road_num**	0	Pseudo Median	-1.00000
ped_cross_hum_con	**M_ped_cross_hum_con**	**IM_ped_cross_hum_con**	0	Pseudo Median	0
ped_cross_phy_facil	**M_ped_cross_phy_facil**	**IM_ped_cross_phy_facil**	0	Pseudo Median	0
light_con	**M_light_con**	**IM_light_con**	0	Pseudo Median	1.00000
weath_con	**M_weath_con**	**IM_weath_con**	0	Pseudo Median	1.00000
road_surf_con	**M_road_surf_con**	**IM_road_surf_con**	0	Pseudo Median	1.00000
spec_con_site	**M_spec_con_site**	**IM_spec_con_site**	0	Pseudo Median	0
carri_haz	**M_carri_haz**	**IM_carri_haz**	0	Pseudo Median	0
urb_or_rur_area	**M_urb_or_rur_area**	**IM_urb_or_rur_area**	0	Pseudo Median	1.00000
did_poli_offi_att	**M_did_poli_offi_att**	**IM_did_poli_offi_att**	0	Pseudo Median	1.00000
tru_road_flag	**M_tru_road_flag**	**IM_tru_road_flag**	0	Pseudo Median	2.00000
hour_of_day	**M_hour_of_day**	**IM_hour_of_day**	0	Pseudo Median	14.00000

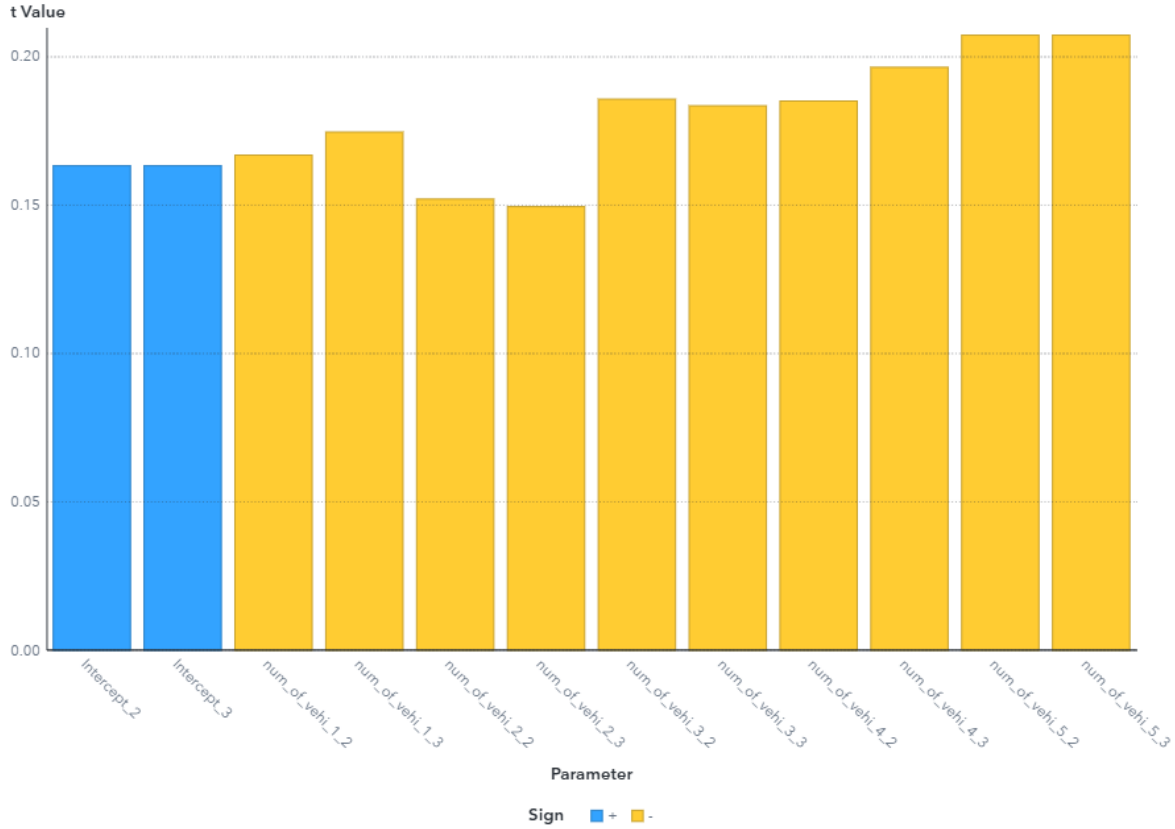
Accident_Fatality_Predict...

"Logistic Regression" Results

by: ta01468@surrey.ac.uk

Contents

t Values by Parameter



This plot displays the absolute value of the t value for each parameter estimate in the logistic regression model. Larger values indicate more significant parameters. The bar that represents the parameter is colored by the sign of the estimate. Bars that are colored as positive (+) correspond to a positive parameter estimate, which indicates an increase in the predicted probability of the target level as the parameter value increases. Bars that are colored as negative (-) correspond to a negative parameter estimate, which indicates a decrease in the predicted probability of the target level as the parameter value increases. The target level to which the parameter estimate corresponds is suffixed to the parameter name (for a cumulative link model, this is only true for the intercept). The most significant parameter is num_of_vehi for the target level "2" with a t value of -0.207.

Parameter Estimates

Effect	Parameter	t Value	Sign
Intercept	Intercept_3	0.1633	+
Intercept	Intercept_2	0.1633	+
num_of_vehi	num_of_vehi_1_3	0.1747	-
num_of_vehi	num_of_vehi_1_2	0.1669	-
num_of_vehi	num_of_vehi_2_3	0.1496	-
num_of_vehi	num_of_vehi_2_2	0.1521	-
num_of_vehi	num_of_vehi_3_3	0.1836	-
num_of_vehi	num_of_vehi_3_2	0.1858	-
num_of_vehi	num_of_vehi_4_3	0.1965	-
num_of_vehi	num_of_vehi_4_2	0.1851	-
num_of_vehi	num_of_vehi_5_3	0.2073	-
num_of_vehi	num_of_vehi_5_2	0.2073	-
num_of_vehi	num_of_vehi_7_3		+
num_of_vehi	num_of_vehi_7_2		+

Estimate	Absolute Estimate	Standard Error	Chi-Square
8.6263	8.6263	52.8104	0.0267
8.6263	8.6263	52.8104	0.0267
-9.2258	9.2258	52.8105	0.0305
-8.8148	8.8148	52.8105	0.0279
-7.8997	7.8997	52.8105	0.0224
-8.0341	8.0341	52.8105	0.0231
-9.6936	9.6936	52.8106	0.0337
-9.8113	9.8113	52.8107	0.0345
-10.3780	10.3780	52.8111	0.0386
-9.7773	9.7773	52.8108	0.0343
-10.9487	10.9487	52.8125	0.0430
-10.9487	10.9487	52.8125	0.0430
0	0		
0	0		

Pr > Chi-Square	Degrees of Freedom	Predicted Outcome
0.8702	1	3
0.8702	1	2
0.8613	1	3
0.8674	1	2
0.8811	1	3
0.8791	1	2
0.8544	1	3
0.8526	1	2
0.8442	1	3
0.8531	1	2
0.8358	1	3
0.8358	1	2
	0	3
	0	2

Selection Summary

Step	Effect Entered	Effect Removed	Number of Effects
0	Intercept		1
1	num_of_vehi		2
2	loc_auth_ons_distr		3
3		loc_auth_ons_distr	2

SBC	Optimal SBC
5,422.2064	0
5,143.4869	0
4,957.5827	0
4,866.5311	1

Regression Fit Statistics

Statistic	Description	Training	Testing
M2LL	-2 Log Likelihood	5,044.0705	2,158.7138
AIC	AIC (smaller is better)	5,068.0705	2,182.7138
AICC	AICC (smaller is better)	5,068.1978	2,183.0132
SBC	SBC (smaller is better)	5,137.7850	2,242.2494
ASE	Average Square Error	0.6162	0.6168

Score Inputs

Name	Role	Variable Level	Type
num_of_vehi	INPUT	NOMINAL	N

Variable Type	Variable Label	Variable Format	Variable Length
double			8

Score Outputs

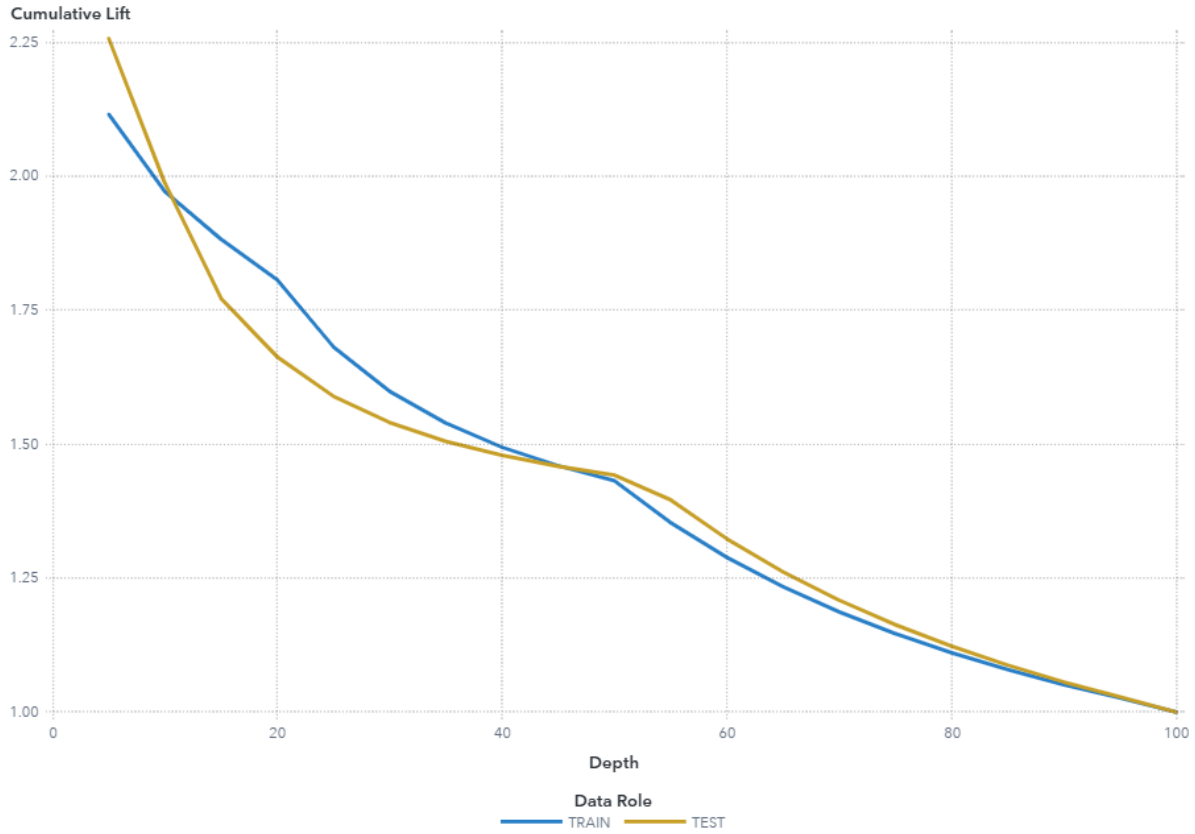
Name	Role	Type	Variable Type
EM_CLASSIFICATION	CLASSIFICATION	C	char
EM_EVENTPROBABILITY	PREDICT	N	double
EM_PROBABILITY	PREDICT	N	double
I_acci_severity	CLASSIFICATION	C	char
P_acci_severity1	PREDICT	N	double
P_acci_severity2	PREDICT	N	double
P_acci_severity3	PREDICT	N	double

Variable Label	Variable Format	Variable Length	Creator
Predicted for acci_severity		12	logisticreg
Probability for acci_severity=1		8	logisticreg
Probability of Classification		8	logisticreg
Into: acci_severity		12	logisticreg
Predicted: acci_severity=1		8	logisticreg
Predicted: acci_severity=2		8	logisticreg
Predicted: acci_severity=3		8	logisticreg

Function	Creator GUID
CLASSIFICATION	637f35cd-d552-4e3f-9abb-4c689142a3aa
PREDICT	637f35cd-d552-4e3f-9abb-4c689142a3aa
PREDICT	637f35cd-d552-4e3f-9abb-4

Function	Creator GUID
	c689142a3aa
CLASSIFICATION	637f35cd-d552-4e3f-9abb-4c689142a3aa
PREDICT	637f35cd-d552-4e3f-9abb-4c689142a3aa
PREDICT	637f35cd-d552-4e3f-9abb-4c689142a3aa
PREDICT	637f35cd-d552-4e3f-9abb-4c689142a3aa

Cumulative Lift



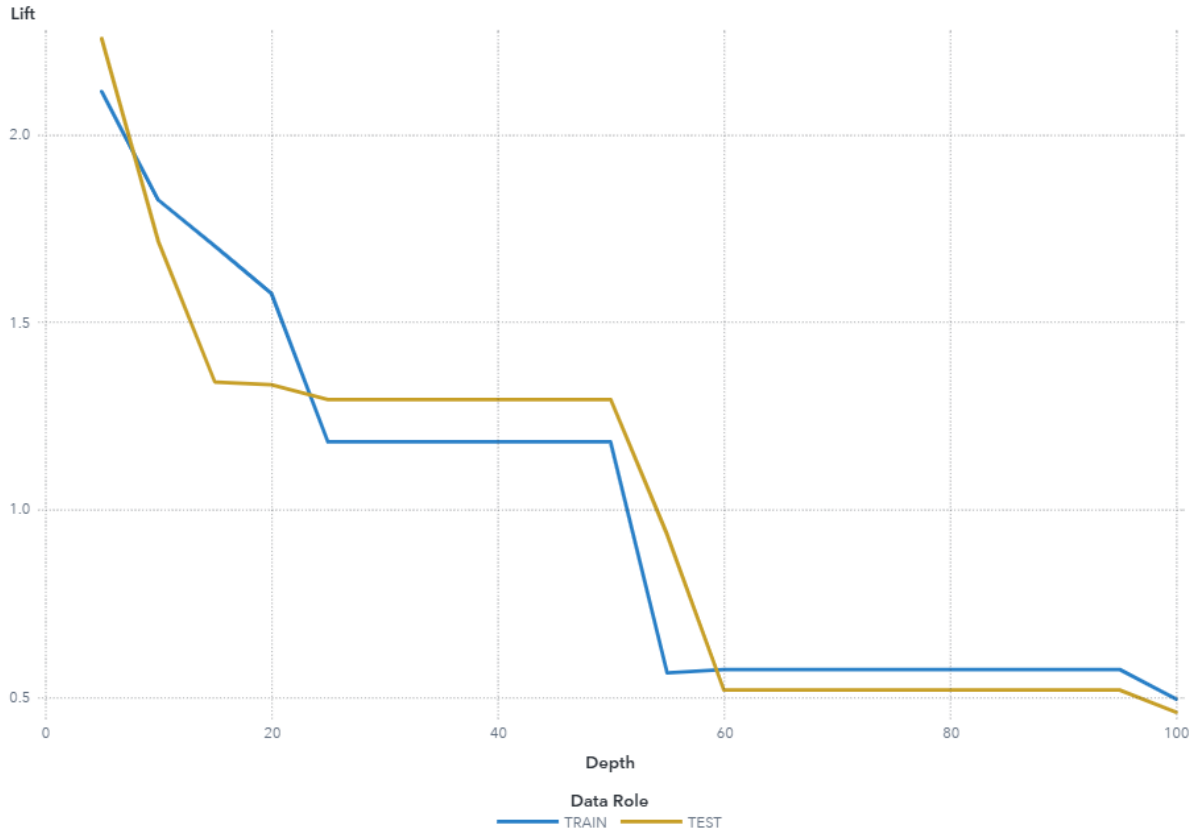
The TRAIN partition has a Cumulative Lift of 1.97 in the 10% quantile (depth of 10) meaning there are 1.97 times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

The TEST partition has a Cumulative Lift of 1.99 in the 10% quantile (depth of 10) meaning there are 1.99 times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

Cumulative lift is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. The cumulative lift for a particular quantile is the ratio of the number of events across all quantiles up to and including the current quantile to the number of events that would be there at random, or equivalently, the ratio of the cumulative response percentage to the baseline response percentage. The cumulative lift at depth 10 includes the top 10% of the data, which is the first 2

quantiles, which would have 10% of the events at random. Thus, cumulative lift measures how much more likely it is to observe an event in the quantiles than by selecting observations at random.

Lift

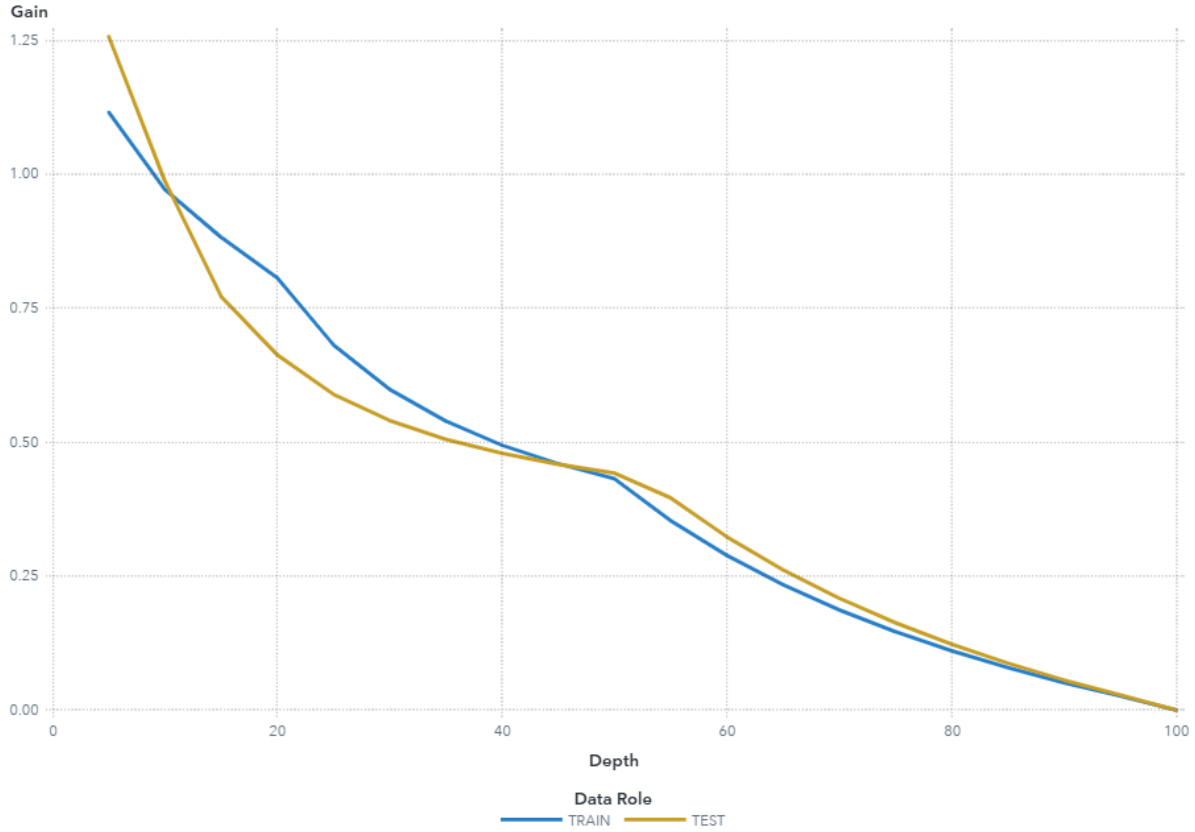


The TRAIN partition has a Lift of 2.12 in the 5% quantile (depth of 5) meaning there are 2.12 times more events in that quantile than expected by random (5% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

The TEST partition has a Lift of 2.26 in the 5% quantile (depth of 5) meaning there are 2.26 times more events in that quantile than expected by random (5% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

Lift is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Lift is the ratio of the number of events in that quantile to the number of events that would be there at random, or equivalently, the ratio of the response percentage to the baseline response percentage. With 20 quantiles, it is expected that 5% of the events occur in each quantile. Thus, Lift measures how much more likely it is to observe an event in each quantile than by selecting observations at random.

Gain

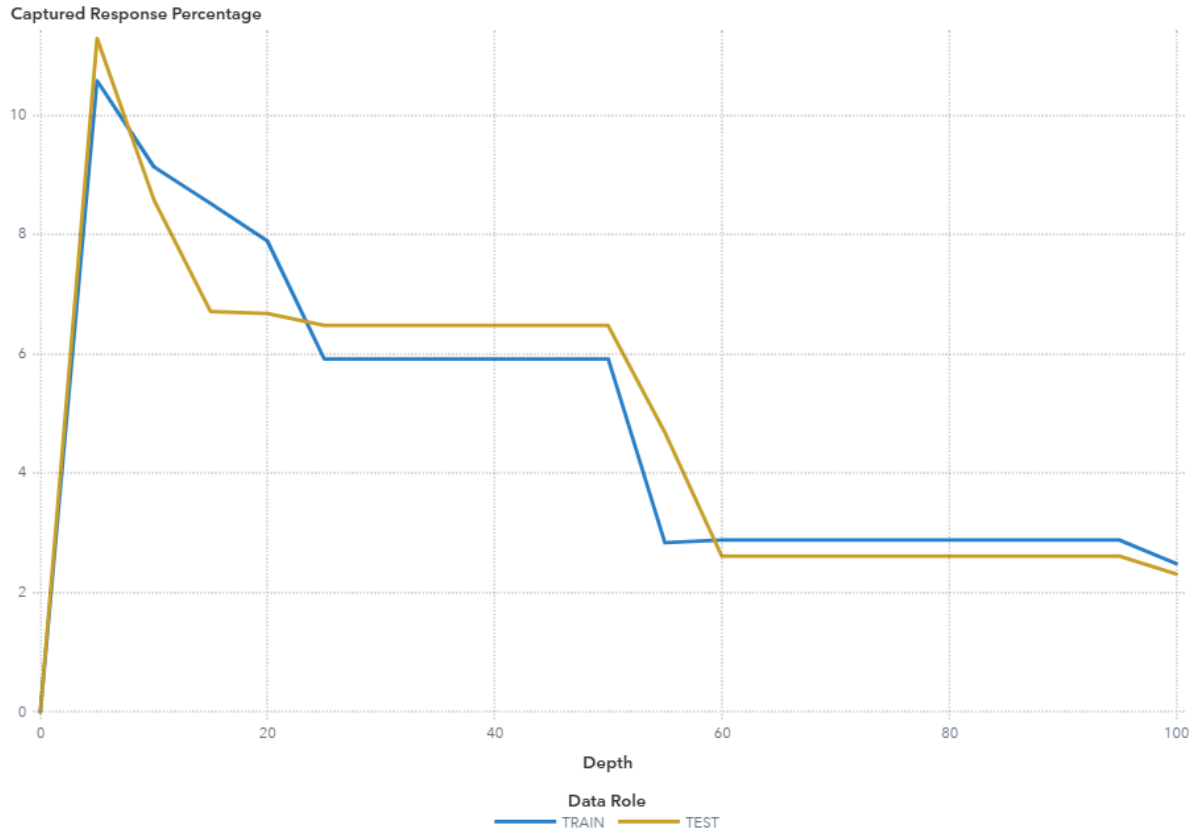


The TRAIN partition has a Gain of 1 at the 10% quantile (depth of 10). Because this value is greater than 0, it is better to use your model to identify responders than no model, based on the selected partition. The best possible value of Gain for this partition at depth 10 is 1.81.

The TEST partition has a Gain of 1 at the 10% quantile (depth of 10). Because this value is greater than 0, it is better to use your model to identify responders than no model, based on the selected partition. The best possible value of Gain for this partition at depth 10 is 1.8.

Gain is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Gain is a cumulative measure for the quantiles up to an including the current one and is calculated as (number of events in the quantiles) / (number of events expected by random) - 1. With 20 quantiles, it is expected that 5% of the events occur in each quantile. Note that the value of Gain is the same as the value of Cumulative Lift - 1. If the value of Gain is greater than 0, then your model is better at identifying events than using no model.

Captured Response Percentage

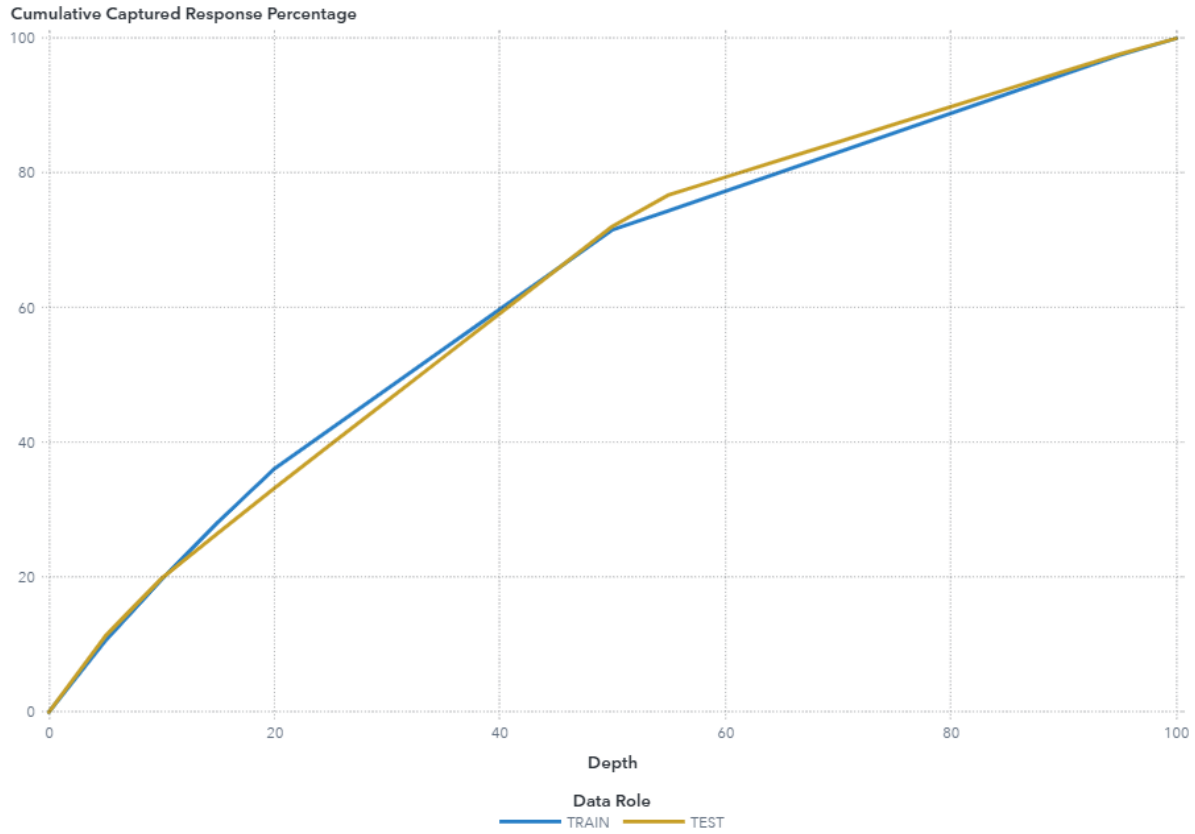


At the 5% quantile (depth of 5), the TRAIN partition has a Captured response percentage of 10.6 (compared to the expected value of 5 for no model). The best possible value of Captured response percentage for this partition at depth 5 is 14.06.

At the 5% quantile (depth of 5), the TEST partition has a Captured response percentage of 11.3 (compared to the expected value of 5 for no model). The best possible value of Captured response percentage for this partition at depth 5 is 14.02.

Captured response percentage is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Captured response percentage is the percentage of the total number of events that are in that quantile. With no model, it is expected that 5% of the events are in each quantile.

Cumulative Captured Response Percentage

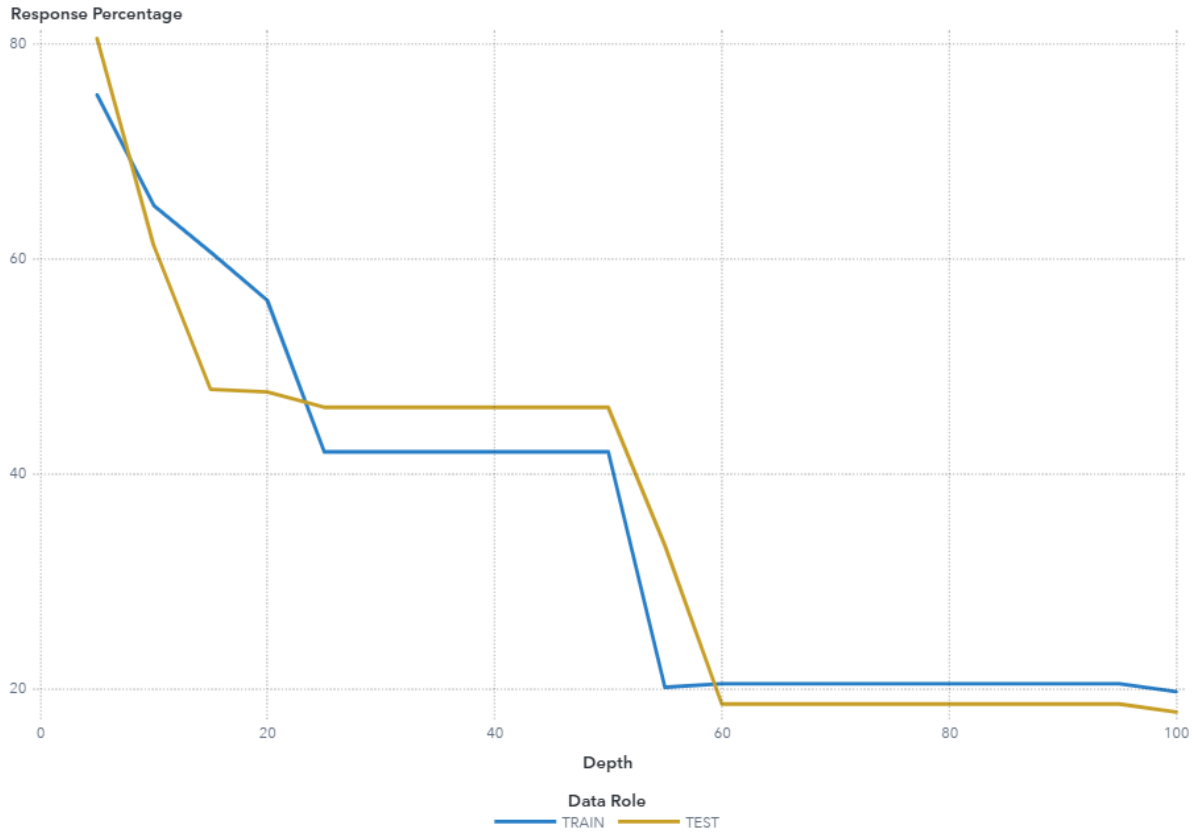


In the top 10% of the data (depth 10), the TRAIN partition has a Cumulative captured response percentage of 19.7 (compared to the expected value of 10 for no model). The best possible value of Cumulative captured response percentage for this partition at depth 10 is 28.12.

In the top 10% of the data (depth 10), the TEST partition has a Cumulative captured response percentage of 19.9 (compared to the expected value of 10 for no model). The best possible value of Cumulative captured response percentage for this partition at depth 10 is 28.04.

Cumulative captured response percentage is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. The cumulative captured response percentage for a particular quantile is the percentage of the total number of events that are in the quantiles up to and including the current quantile. With no model, it is expected that 5% of the events are in each quantile, so the cumulative captured response percentage at depth 10 would be 10%.

Response Percentage

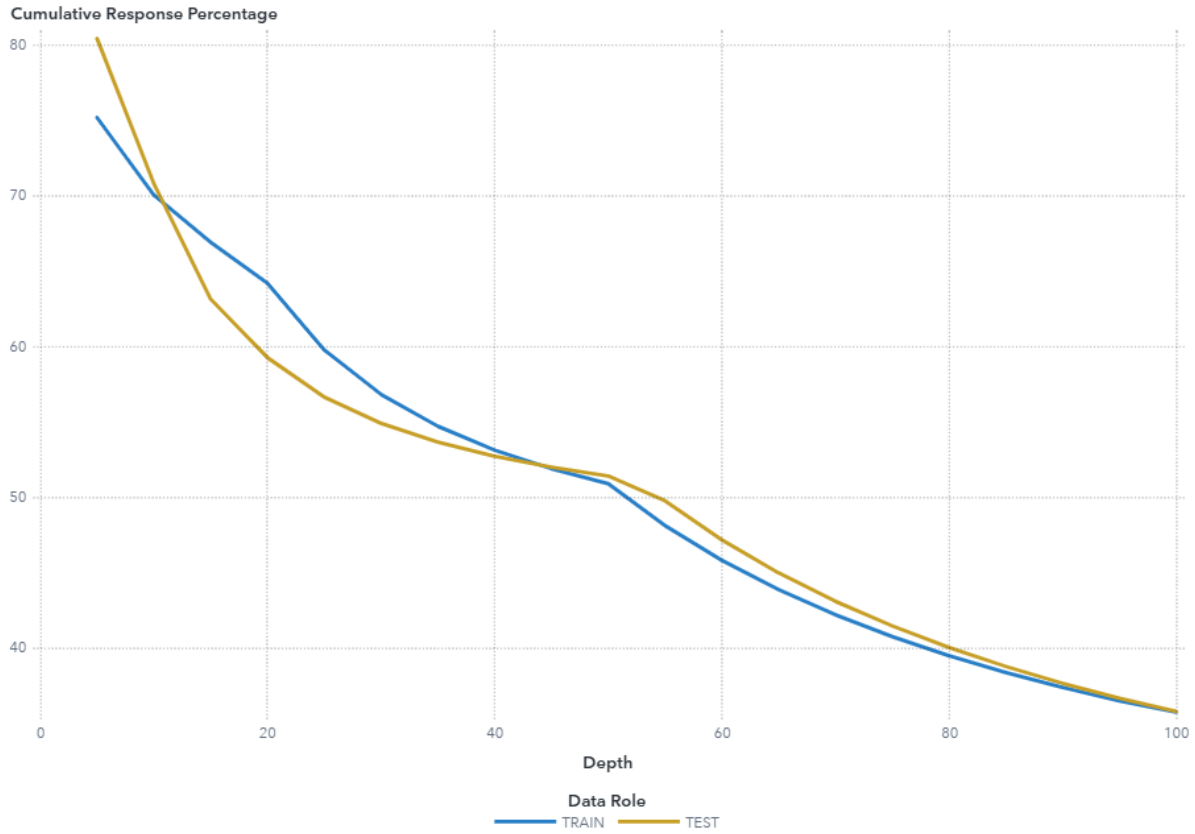


At the 5% quantile (depth of 5), the TRAIN partition has a Response percentage of 75.2. The best possible value of Response percentage for this partition at depth 5 is 100.

At the 5% quantile (depth of 5), the TEST partition has a Response percentage of 80.5. The best possible value of Response percentage for this partition at depth 5 is 100.

Response percentage is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Response percentage is the percentage of observations that are events in that quantile. With no model, it is expected that the response percentage is constant across quantiles, 100*overall-event-rate. This is also called the baseline response percentage.

Cumulative Response Percentage

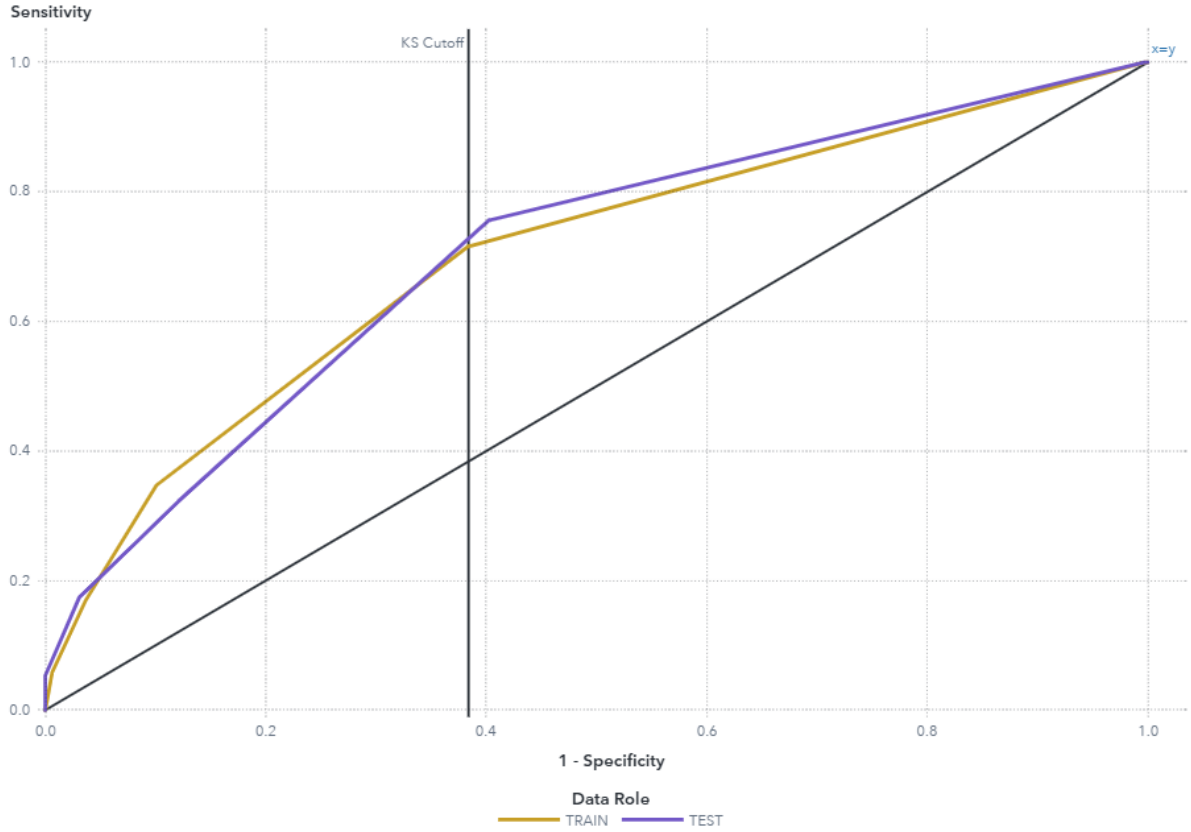


In the top 10% of the data (depth 10), the TRAIN partition has a Cumulative response percentage of 70.1. The best possible value of Cumulative response percentage for this partition at depth 10 is 100.

In the top 10% of the data (depth 10), the TEST partition has a Cumulative response percentage of 70.8. The best possible value of Cumulative response percentage for this partition at depth 10 is 100.

Cumulative response percentage is calculated by sorting in descending order each partition of the data by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. The cumulative response percentage for a particular quantile is the percentage of observations that are events in the quantiles up to and including the current quantile. With no model, it is expected that the response percentage is constant across quantiles, 100*overall-event-rate. This is also called the baseline response percentage.

ROC



The ROC curve is a plot of sensitivity (the true positive rate) against 1-specificity (the false positive rate), which are both measures of classification based on the confusion matrix. These measures are calculated at various cutoff values. To help identify the best cutoff to use when scoring your data, the KS Cutoff reference line is drawn at the value of 1-specificity where the greatest difference between sensitivity and 1-specificity is observed for the TRAIN partition. The KS Cutoff line is drawn at the cutoff value 0.36, where the 1-specificity value is 0.384 and the sensitivity value is 0.717.

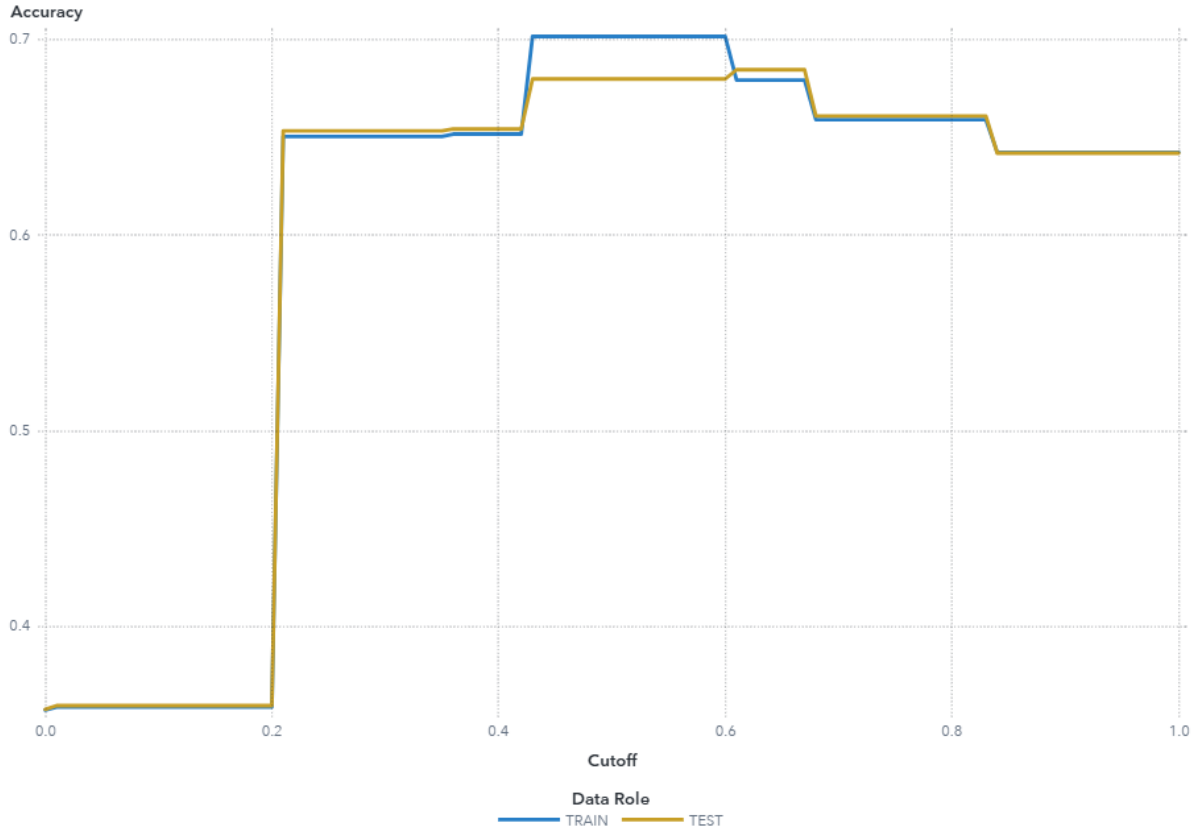
Cutoff values range from 0 to 1, inclusive, in increments of 0.01. At each cutoff value, the predicted target classification is determined by whether P_acci_severity1, which is the predicted probability of the event "1" for the target acci_severity, is greater than or equal to the cutoff value. When P_acci_severity1 is greater than or equal to the cutoff value, then the predicted classification is the event, otherwise it is a non-event.

The confusion matrix for each cutoff value contains four cells that display the true positives for events that are correctly classified (TP), false positives for non-events that are classified as events (FP), false negatives for events that are classified as non-events (FN), and true negatives for non-events that are classified as non-events (TN). True negatives include non-event classifications that specify a different non-event. Sensitivity is calculated as TP / (TP + FN). Specificity, the true negative rate,

is calculated as TN / (TN + FP), so 1-specificity is FP / (TN + FP). The values of sensitivity and 1-specificity are plotted at each cutoff value.

A ROC curve that rapidly approaches the upper-left corner of the graph, where the difference between sensitivity and 1-specificity is the greatest, indicates a more accurate model. A diagonal line where sensitivity = 1-specificity indicates a random model.

Accuracy

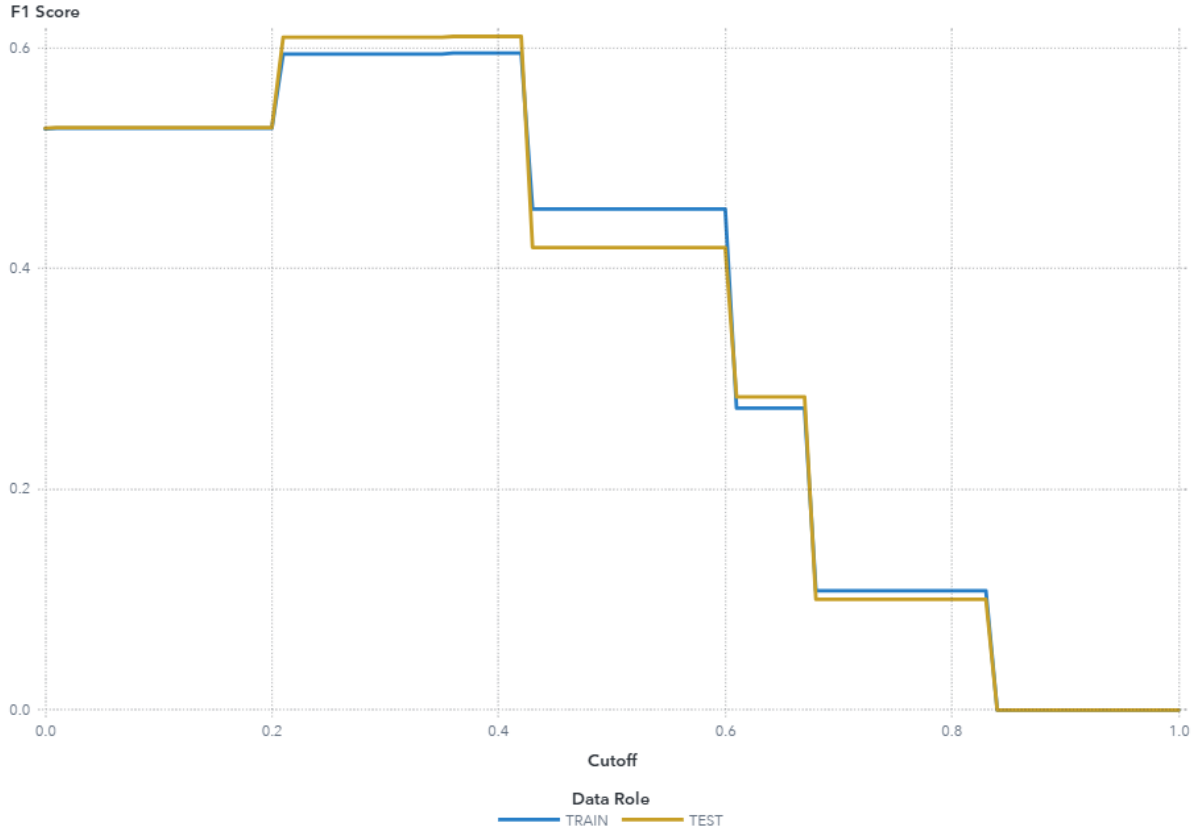


For this model, the accuracy in the TEST partition at the cutoff of 0.5 is 0.68.

For this model, the accuracy in the TRAIN partition at the cutoff of 0.5 is 0.702.

Accuracy is the proportion of observations that are correctly classified as either an event or non-event, calculated at various cutoff values. Cutoff values range from 0 to 1, inclusive, in increments of 0.01. At each cutoff value, the predicted target classification is determined by whether P_acci_severity1, which is the predicted probability of the event "1" for the target acci_severity, is greater than or equal to the cutoff value. When P_acci_severity1 is greater than or equal to the cutoff value, then the predicted classification is the event, otherwise it is a non-event. When the predicted classification and the actual classification are both events (true positives) or both non-events (true negatives), the observation is correctly classified. If the predicted classification and actual classification disagree, then the observation is incorrectly classified. Accuracy is calculated as (true positives + true negatives) / (total observations).

F1 Score



For this model, the F1 score in the TEST partition at the cutoff of 0.5 is 0.419.

For this model, the F1 score in the TRAIN partition at the cutoff of 0.5 is 0.454.

The F1 score combines the measures of precision and recall (or sensitivity), which are measures of classification based on the confusion matrix that are calculated at various cutoff values. Cutoff values range from 0 to 1, inclusive, in increments of 0.01. At each cutoff value, the predicted target classification is determined by whether P_acci_severity1, which is the predicted probability of the event "1" for the target acci_severity, is greater than or equal to the cutoff value. When P_acci_severity1 is greater than or equal to the cutoff value, then the predicted classification is the event, otherwise it is a non-event.

The confusion matrix for each cutoff value contains four cells that display the true positives for events that are correctly classified (TP), false positives for non-events that are classified as events (FP), false negatives for events that are classified as non-events (FN), and true negatives for non-events that are classified as non-events (TN). True negatives include non-event classifications that specify a different non-event.

Precision is calculated as TP / (TP + FP), and recall (or sensitivity) is calculated as

TP / (TP + FN). The F1 score is calculated as 2*Precision*Recall / (Precision + Recall), which is the harmonic mean of Precision and Recall. Larger F1 scores indicate a more accurate model.

Fit Statistics

Target Name	Data Role	Partition Indicator	Formatted Partition
acci_severity	TEST	2	2
acci_severity	TRAIN	1	1

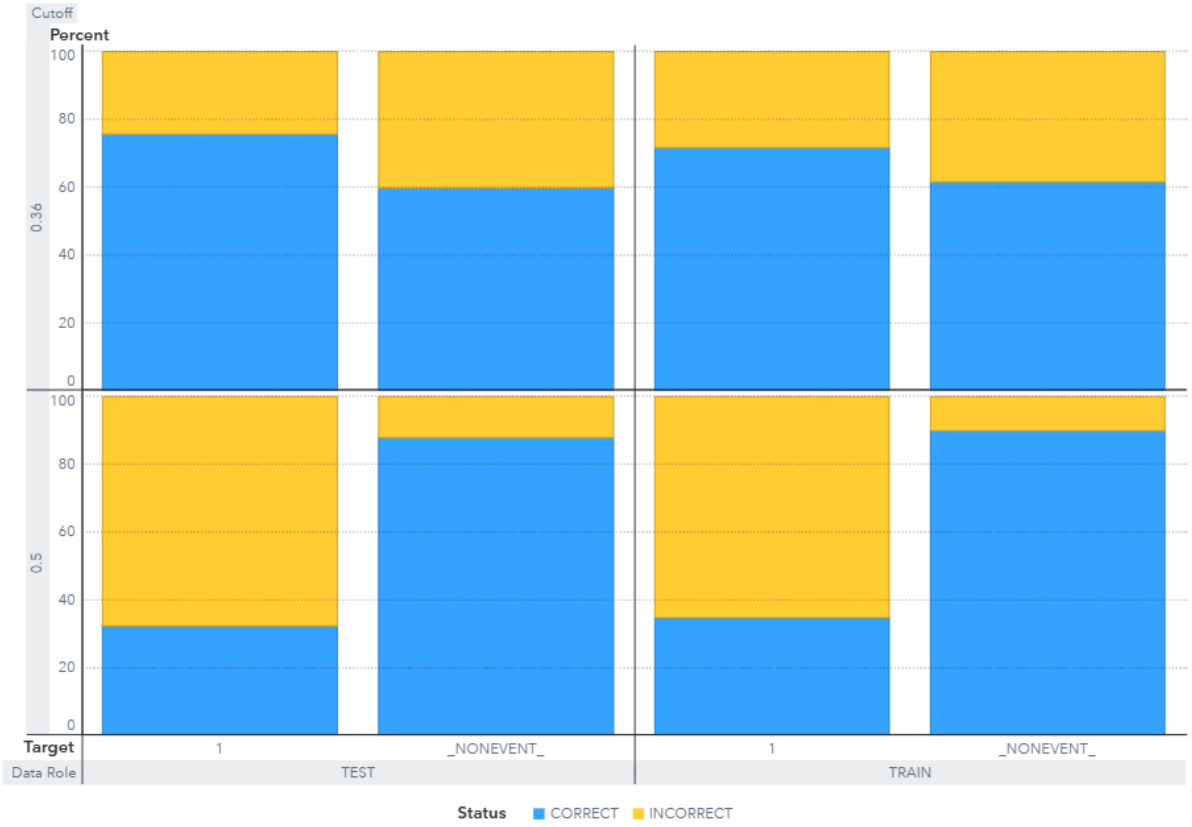
Number of Observations	Average Squared Error	Divisor for ASE	Root Average Squared Error
1,056	0.2056	1,056	0.4534
2,467	0.2054	2,467	0.4532

Misclassification Rate	Multi-Class Log Loss	KS (Youden)	Area Under ROC
0.5275	1.0232	0.3540	0.7028
0.5334	1.0237	0.3323	0.6994

Gini Coefficient	Gamma	Tau	KS Cutoff
0.4055	0.5661	0.1866	0.3600
0.3989	0.5643	0.1833	0.3600

KS at Default Cutoff	Misclassification Rate at KS Cutoff (Event)	Misclassification Rate (Event)
0.2018	0.3456	0.3201
0.2460	0.3482	0.2983

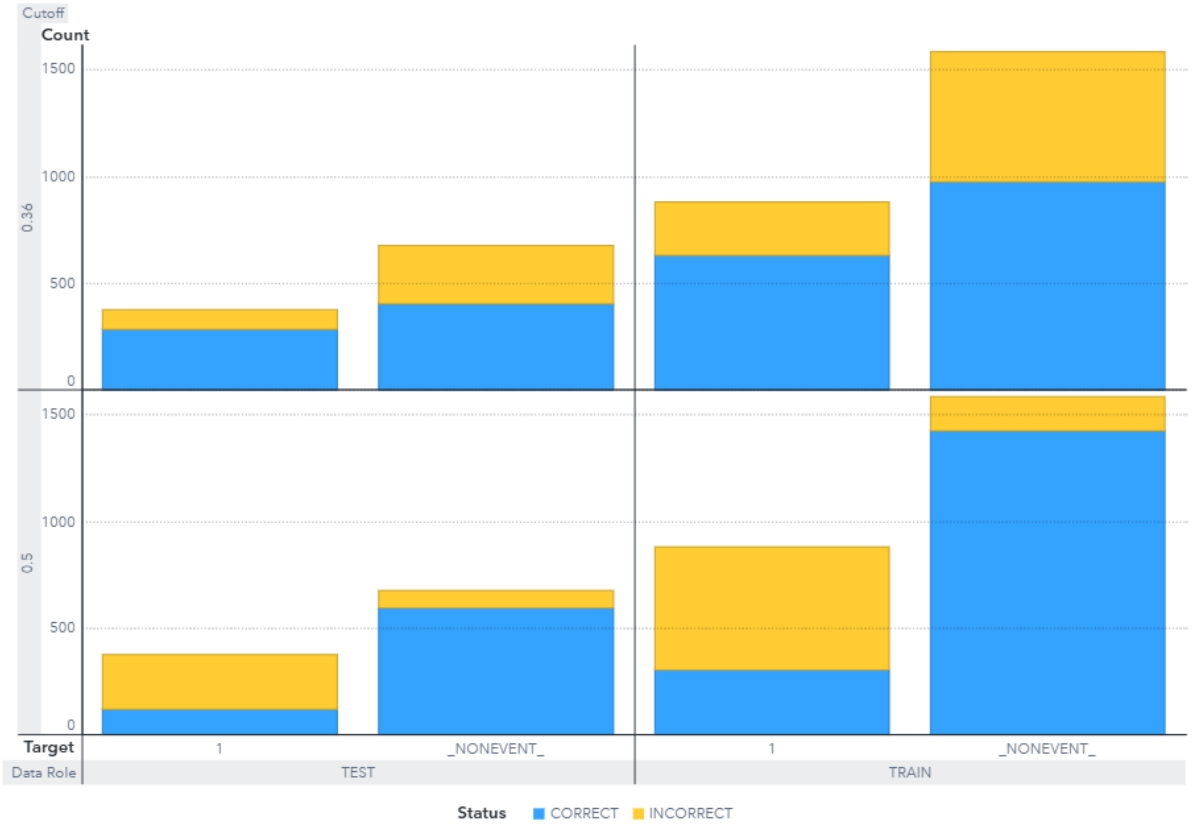
Percentage Plot



The Event Classification report is a visual representation of the confusion matrix at various cutoff values for each partition. The classification cutoffs used in the plot are the default (0.5) and these KS cutoff values for existing partitions: 0.36 (TRAIN), 0.36 (TEST).

For this data, for the bar corresponding to the event level of acci_severity, " 1", the segment of the bar colored as "CORRECT" corresponds to true positives.

Count Plot



The Event Classification report is a visual representation of the confusion matrix at various cutoff values for each partition. The classification cutoffs used in the plot are the default (0.5) and these KS cutoff values for existing partitions: 0.36 (TRAIN), 0.36 (TEST).

For this data, for the bar corresponding to the event level of acci_severity, " 1", the segment of the bar colored as "CORRECT" corresponds to true positives.

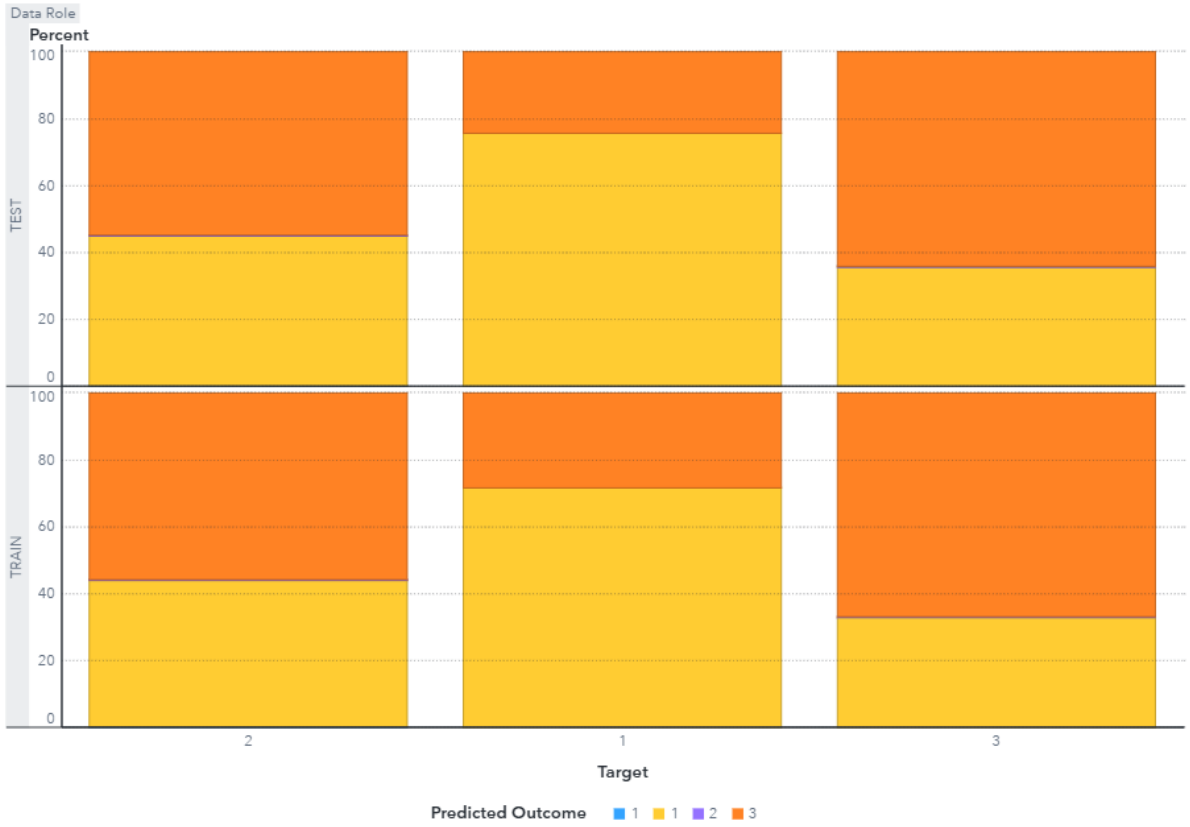
Table

Cutoff	Cutoff Source	Target Name	Response
0.3600	KS	acci_severity	CORRECT
0.3600	KS	acci_severity	INCORRECT
0.3600	KS	acci_severity	CORRECT
0.3600	KS	acci_severity	INCORRECT
0.5000	Default	acci_severity	CORRECT
0.5000	Default	acci_severity	INCORRECT
0.5000	Default	acci_severity	CORRECT
0.5000	Default	acci_severity	INCORRECT

Event	Value	Training Frequency	Validation Frequency
1	True Positive	632	
1	False Negative	250	
NONEVENT	True Negative	976	
NONEVENT	False Positive	609	
1	True Positive	306	
1	False Negative	576	
NONEVENT	True Negative	1,425	
NONEVENT	False Positive	160	

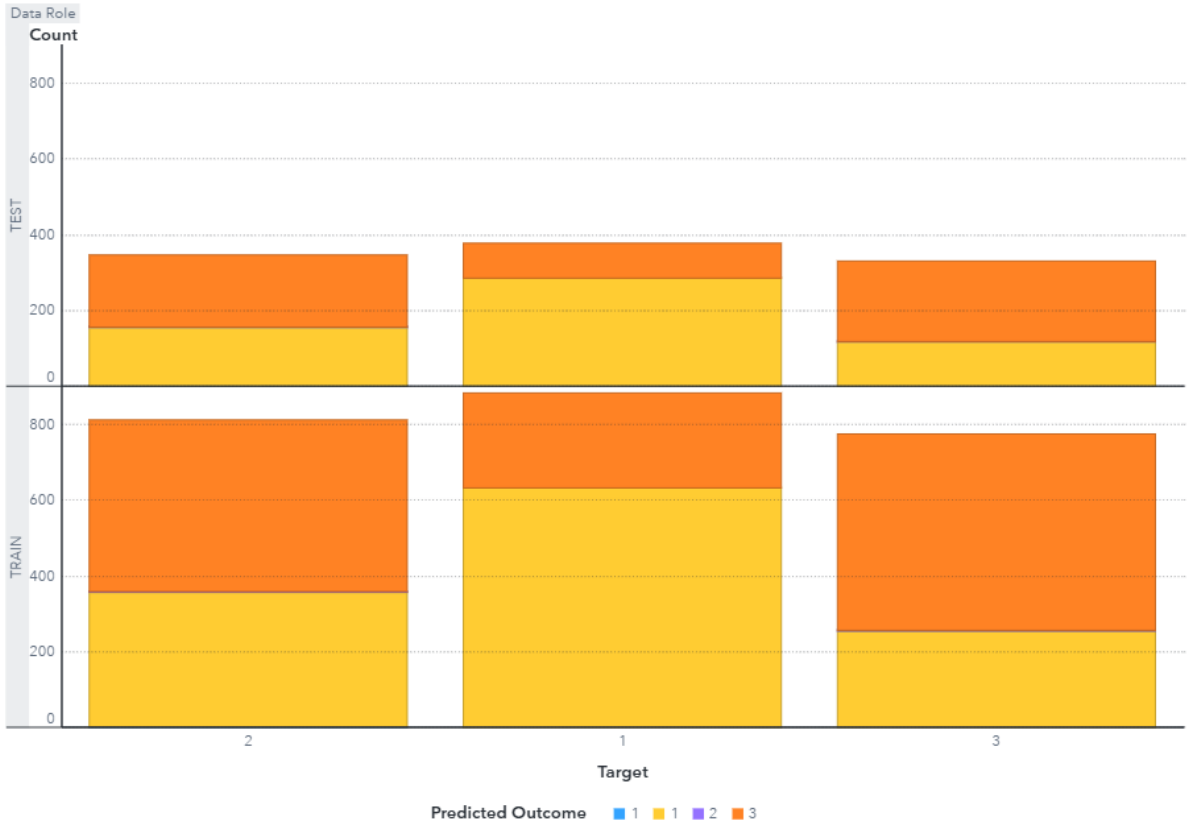
Test Frequency	Training Percentage	Validation Percentage	Test Percentage
286	71.6553		75.6614
92	28.3447		24.3386
405	61.5773		59.7345
273	38.4227		40.2655
122	34.6939		32.2751
256	65.3061		67.7249
596	89.9054		87.9056
82	10.0946		12.0944

Percentage Plot



The Nominal Classification report displays either the percentage of or the number of observations predicting each target level. The plot is segmented by target level and partition level. The target level with the greatest predicted probability is the predicted outcome. A greater number of observations where the target and predicted outcome are the same indicates a better model.

Count Plot



The Nominal Classification report displays either the percentage of or the number of observations predicting each target level. The plot is segmented by target level and partition level. The target level with the greatest predicted probability is the predicted outcome. A greater number of observations where the target and predicted outcome are the same indicates a better model.

Table

Target Name	Data Role	Target	Unformatted Target
acci_severity	TRAIN	1	1
acci_severity	TRAIN	1	1
acci_severity	TRAIN	2	2
acci_severity	TRAIN	2	2
acci_severity	TRAIN	2	2
acci_severity	TRAIN	2	2
acci_severity	TRAIN	3	3
acci_severity	TRAIN	3	3
acci_severity	TRAIN	3	3
acci_severity	TEST	1	1
acci_severity	TEST	1	1
acci_severity	TEST	2	2
acci_severity	TEST	2	2
acci_severity	TEST	2	2
acci_severity	TEST	2	2
acci_severity	TEST	3	3
acci_severity	TEST	3	3
acci_severity	TEST	3	3

Predicted Outcome	Count	Percent	Status
1	632	71.6553	CORRECT
3	250	28.3447	INCORRECT
1	3	0.3699	INCORRECT
1	354	43.6498	INCORRECT
2	2	0.2466	CORRECT
3	452	55.7337	INCORRECT
1	255	32.9457	INCORRECT
2	2	0.2584	INCORRECT
3	517	66.7959	CORRECT

Predicted Outcome	Count	Percent	Status
1	286	75.6614	CORRECT
3	92	24.3386	INCORRECT
1	1	0.2882	INCORRECT
1	155	44.6686	INCORRECT
2	1	0.2882	CORRECT
3	190	54.7550	INCORRECT
1	118	35.6495	INCORRECT
2	1	0.3021	INCORRECT
3	212	64.0483	CORRECT

Properties

Property Name	Property Value
binaryProbCutoff	0.5000
chooseCriterion	SBC
classCoding	GLM
classOrder	FMTASC
codeLocation	mlearning
dataMiningVersion	V2024.03
exactPctlLift	true
explainFidelity	false
explainInfo	false
factorInteractions	false
factorSplit	false
fullDatasetReconstitution	false
hierarchy	NONE
icePlots	false
informativeMiss	false
linkFunction	LOGIT
maxEffects	0
maxNumShapVars	20
maxSteps	0
minEffects	0
missAsLvl	false
nBins	50
nomlinkFunction	GLOGIT
normalize	true
pdNumImportantInputs	5
pdObsSamples	1,000
pdPlots	false
performKernelShap	false

Property Name	Property Value
performLime	false
performVI	false
polynomialDegree	2
reportingOnly	false
seedId	12,345
selectCriterion	SBC
selectMethod	STEPWISE
slEntry	0.0500
slStay	0.0500
specifyRows	RANDOM
stopCriterion	SBC
suppressIntercept	false
tech	NRRIDG
templateRevision	2
train	true
truncateLl	5
truncateUl	95
usePolynomial	false
useSpline	false
useSplineSplit	false
userProbCutoff	false

Output

The SAS System

The GENSELECT Procedure

Model Information	
Data Source	DM_AF3Z7K9DW76JQQTEWKB0AC4J2
Response Variable	acci_severity
Number of Response Levels	3
Distribution	Multinomial
Link Type	Generalized
Link Function	Logit
Optimization Technique	Newton-Raphson with Ridging
Predicted Response Level	I_acci_severity

Number of Observations			
Description	Total	Training	Testing
Number of Observations Read	3523	2467	1056
Number of Observations Used	3519	2464	1055

Response Profile				
Ordered Value	acci_severity	Total Frequency	Training	Testing
1	3	1105	774	331
2	2	1154	808	346
3	1	1260	882	378

Probabilities modeled use acci_severity = 1 as the reference category.

Class Level Information		
Class	Levels	Values
carri_haz	5	0 1 2 3 7
first_road_class	5	1 3 4 5 6
junc_detail	9	0 1 2 3 5 6 7 8 9
loc_auth_ons_distr	11	E07000207 E07000208 E07000209 E07000210 E07000211 E07000212 E07000213 E07000214 E07000215 E07000216 E07000217
num_of_casu	8	1 2 3 4 5 6 7 9
num_of_vehi	6	1 2 3 4 5 7
ped_cross_hum_con	3	0 1 2
ped_cross_phy_facil	6	0 1 4 5 7 8
road_type	6	1 2 3 6 7 9
spec_con_site	6	0 1 3 4 5 7
speed_limit	6	20 30 40 50 60 70
time_category	4	after eveni morni night
urb_or_rur_area	2	1 2
weath_con	8	1 2 3 4 5 7 8 9

Selection Information	
Selection Method	Stepwise
Select Criterion	SBC
Choose Criterion	SBC
Stop Criterion	SBC
Effect Hierarchy Enforced	None
Stop Horizon	3

Selection Details

Convergence criterion (ABSGCONV=1E-7) satisfied.

Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	SBC
0	Intercept		1	5422.2064
1	num_of_vehi		2	5143.4869
2	loc_auth_ons_distr		3	4957.5827
3		loc_auth_ons_distr	2	4866.5311*

* Optimal Value Of Criterion

Stepwise selection stopped because adding or removing an effect does not improve the SBC criterion.

The model at step 3 is selected where SBC is 4866.531.

Selected Effects: Intercept num_of_vehi

Selected Model

Dimensions	
Columns in Design	14
Number of Effects	2
Max Effect Columns	12
Rank of Design	12
Parameters in Optimization	12

Fit Statistics		
Description	Training	Testing
-2 Log Likelihood	5044.07049	2158.71380
AIC (smaller is better)	5068.07049	2182.71380
AICC (smaller is better)	5068.19778	2183.01323
SBC (smaller is better)	5137.78498	2242.24936
Average Square Error	0.61615	0.61676

Parameter Estimates						
Parameter	acci_severity	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	3	1	8.626295	52.810402	0.0267	0.8702
Intercept	2	1	8.626295	52.810402	0.0267	0.8702
num_of_vehi 1	3	1	-9.225806	52.810484	0.0305	0.8613
num_of_vehi 1	2	1	-8.814770	52.810466	0.0279	0.8674
num_of_vehi 2	3	1	-7.899713	52.810458	0.0224	0.8811
num_of_vehi 2	2	1	-8.034073	52.810461	0.0231	0.8791
num_of_vehi 3	3	1	-9.693556	52.810638	0.0337	0.8544
num_of_vehi 3	2	1	-9.811339	52.810660	0.0345	0.8526
num_of_vehi 4	3	1	-10.378049	52.811056	0.0386	0.8442
num_of_vehi 4	2	1	-9.777275	52.810804	0.0343	0.8531
num_of_vehi 5	3	1	-10.948682	52.812481	0.0430	0.8358
num_of_vehi 5	2	1	-10.948682	52.812481	0.0430	0.8358
num_of_vehi 7	3	0	0	.	.	.
num_of_vehi 7	2	0	0	.	.	.

Score Code Variables for Predicted Probability	
acci_severity	Variable
1	P_acci_severity1
3	P_acci_severity3
2	P_acci_severity2

Task Timing		
Task	Seconds	Percent
Setup and Parsing	0.01	1.79%
Levelization	0.01	2.40%
Model Initialization	0.00	1.14%
SSCP Computation	0.01	2.64%
Model Selection	0.33	91.54%
Producing Score Code	0.00	0.34%
Display	0.00	0.10%
Cleanup	0.00	0.00%
Total	0.36	100.00%

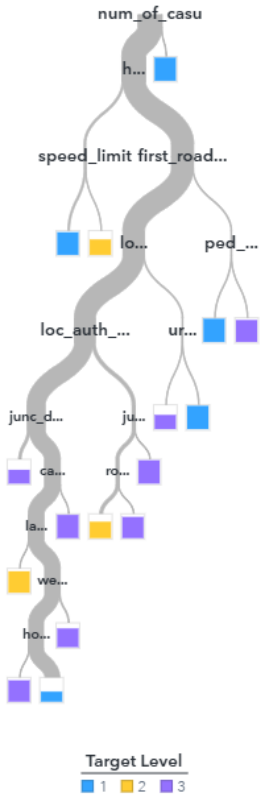
Accident_Fatality_Predict...

"Decision Tree" Results

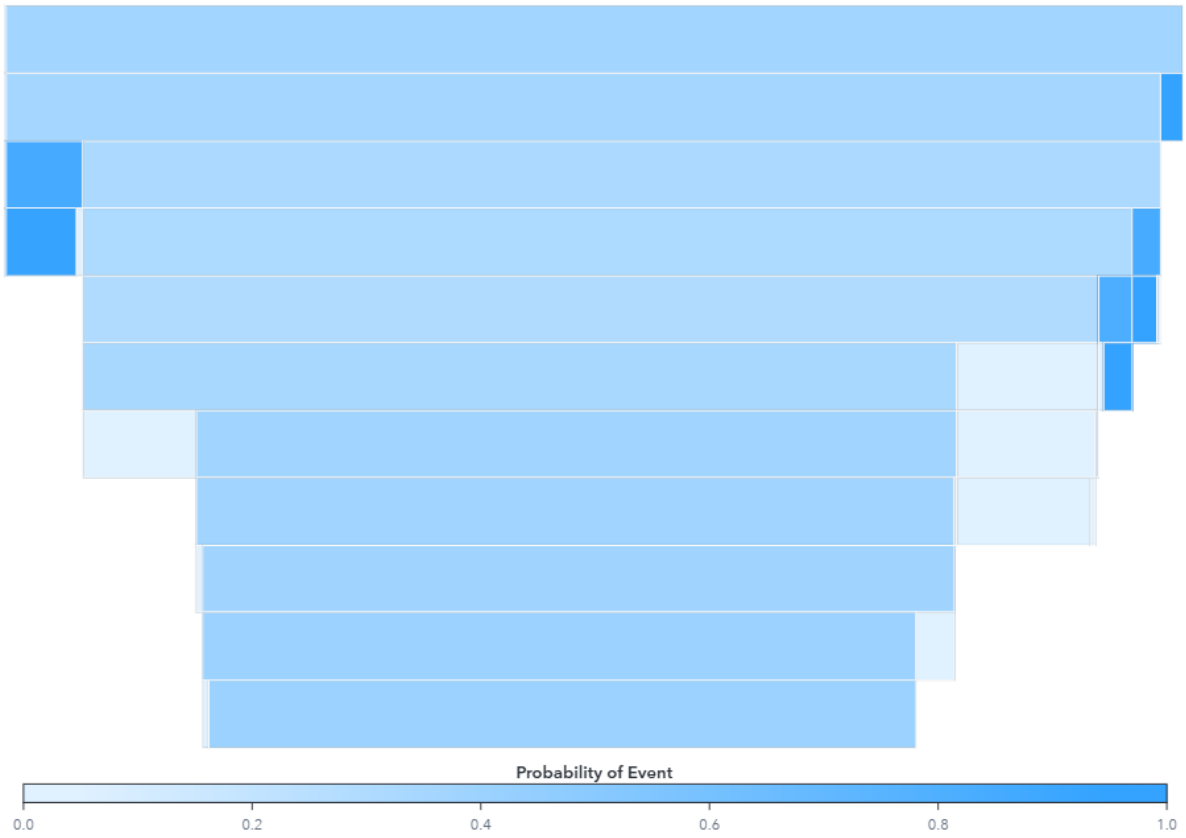
by: ta01468@surrey.ac.uk

Contents

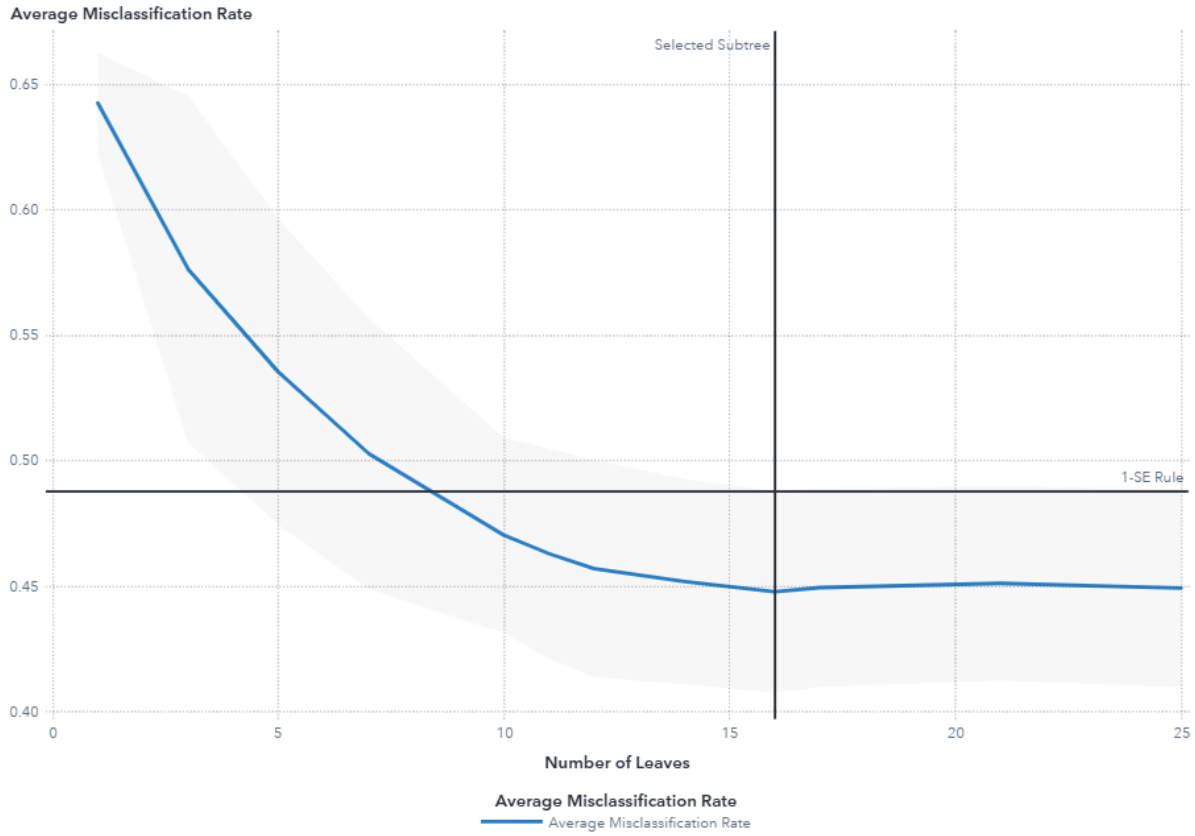
Tree Diagram



Treemap



Cross Validation Cost-Complexity



This plot shows how the average of the misclassification rate across folds changes for subtrees, which are created by cost-complexity pruning of the full decision tree to various numbers of leaves based on cross validation. The band around the line ranges from the average misclassification rate minus one standard error (SE) to the average misclassification rate plus one SE. The reference line for the 1-SE Rule occurs at the value of 0.488, the minimum average misclassification rate plus one SE. When the property for the 1-SE rule is selected, the smallest subtree for which the average misclassification rate is less than this value is used; otherwise, the subtree with the minimum average misclassification rate is used. For this decision tree model, the selected subtree has 16 leaves with an average misclassification rate across folds of 0.448.

Variable Importance

Variable Name	Training Relative Importance	Count	Training Importance
hour_of_day	1	2	85.4301
junc_detail	0.5337	2	45.5964
loc_auth_ons_distr	0.5146	1	43.9653
longitude	0.4049	1	34.5935
first_road_num	0.3553	1	30.3521
num_of_casu	0.3323	1	28.3856
weath_con	0.3105	1	26.5254
speed_limit	0.2329	1	19.9007
urb_or_rur_area	0.1641	1	14.0177
ped_cross_hum_con	0.1267	1	10.8281
latitude	0.0854	1	7.2924
road_type	0.0825	1	7.0448
carri_haz	0.0408	1	3.4867

Score Inputs

Name	Role	Variable Level	Type
carri_haz	INPUT	NOMINAL	N
first_road_num	INPUT	INTERVAL	N
hour_of_day	INPUT	INTERVAL	N
junc_detail	INPUT	NOMINAL	N
latitude	INPUT	INTERVAL	N
loc_auth_ons_distr	INPUT	NOMINAL	C
longitude	INPUT	INTERVAL	N
num_of_casu	INPUT	NOMINAL	N
ped_cross_hum_con	INPUT	NOMINAL	N
road_type	INPUT	NOMINAL	N
speed_limit	INPUT	NOMINAL	N
urb_or_rur_area	INPUT	BINARY	N
weath_con	INPUT	NOMINAL	N

Variable Type	Variable Label	Variable Format	Variable Length
double			8
double			8
double			8
double			8
double			8
varchar			9
double			8
double			8
double			8
double			8
double			8
double			8
double			8

Score Outputs

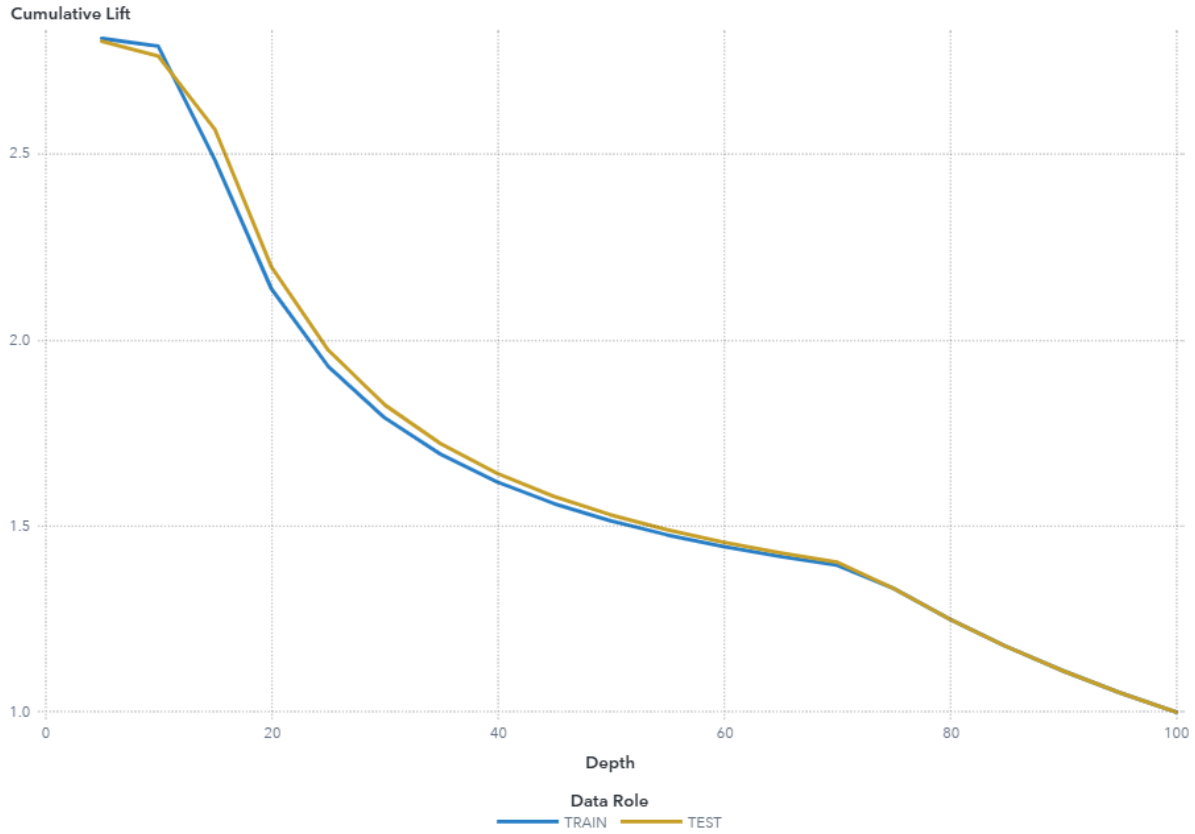
Name	Role	Type	Variable Type
EM_CLASSIFICATION	CLASSIFICATION	C	char
EM_EVENTPROBABILITY	PREDICT	N	double
EM_PROBABILITY	PREDICT	N	double
I_acci_severity	CLASSIFICATION	C	char
P_acci_severity1	PREDICT	N	double
P_acci_severity2	PREDICT	N	double
P_acci_severity3	PREDICT	N	double
WARN	ASSESS	C	char

Variable Label	Variable Format	Variable Length	Creator
Predicted for acci_severity		12	tree
Probability for acci_severity=1		8	tree
Probability of Classification		8	tree
Into: acci_severity		32	tree
Predicted: acci_severity=1		8	tree
Predicted: acci_severity=2		8	tree
Predicted: acci_severity=3		8	tree
Warnings		4	tree

Function	Creator GUID
CLASSIFICATION	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3
PREDICT	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3

Function	Creator GUID
PREDICT	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3
CLASSIFICATION	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3
PREDICT	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3
PREDICT	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3
PREDICT	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3
ASSESS	8c30316c-3ed3-4883-8c3a-f5ffe1867aa3

Cumulative Lift



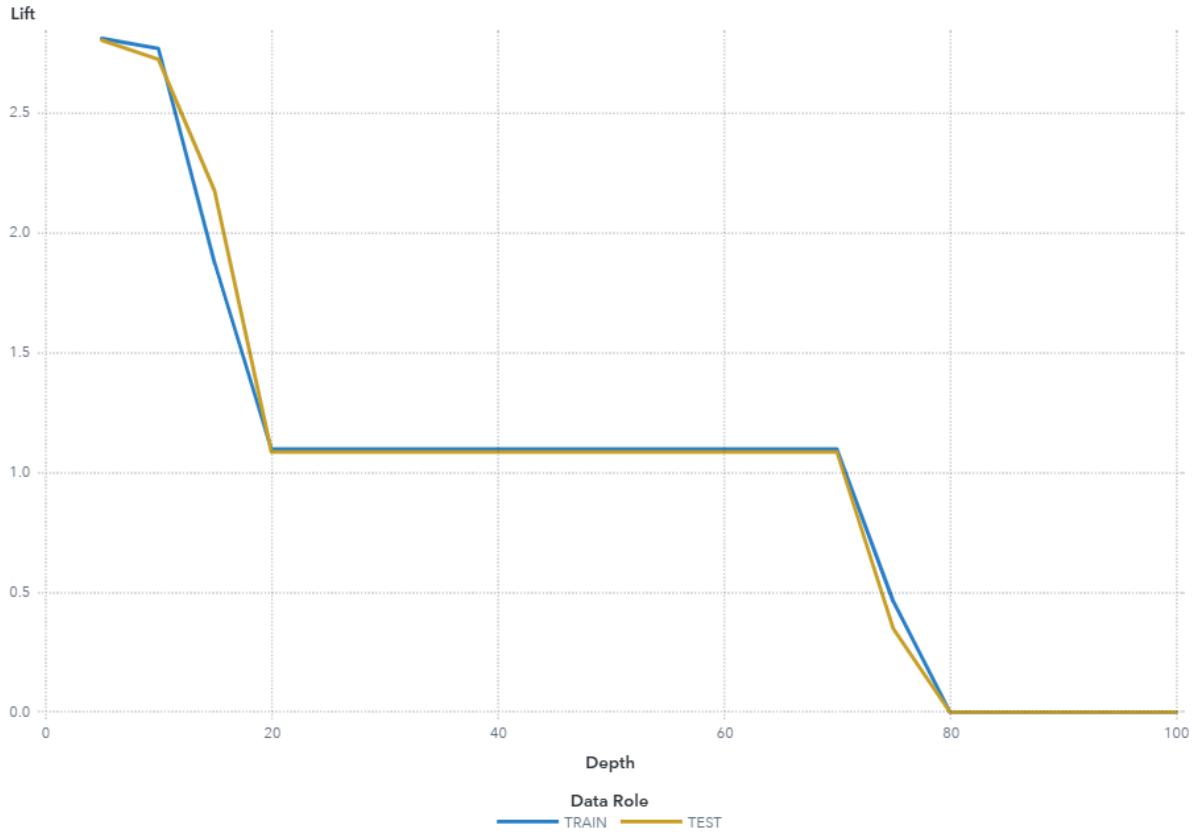
The TRAIN partition has a Cumulative Lift of 2.79 in the 10% quantile (depth of 10) meaning there are 2.79 times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

The TEST partition has a Cumulative Lift of 2.76 in the 10% quantile (depth of 10) meaning there are 2.76 times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

Cumulative lift is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. The cumulative lift for a particular quantile is the ratio of the number of events across all quantiles up to and including the current quantile to the number of events that would be there at random, or equivalently, the ratio of the cumulative response percentage to the baseline response percentage. The cumulative lift at depth 10 includes the top 10% of the data, which is the first 2

quantiles, which would have 10% of the events at random. Thus, cumulative lift measures how much more likely it is to observe an event in the quantiles than by selecting observations at random.

Lift

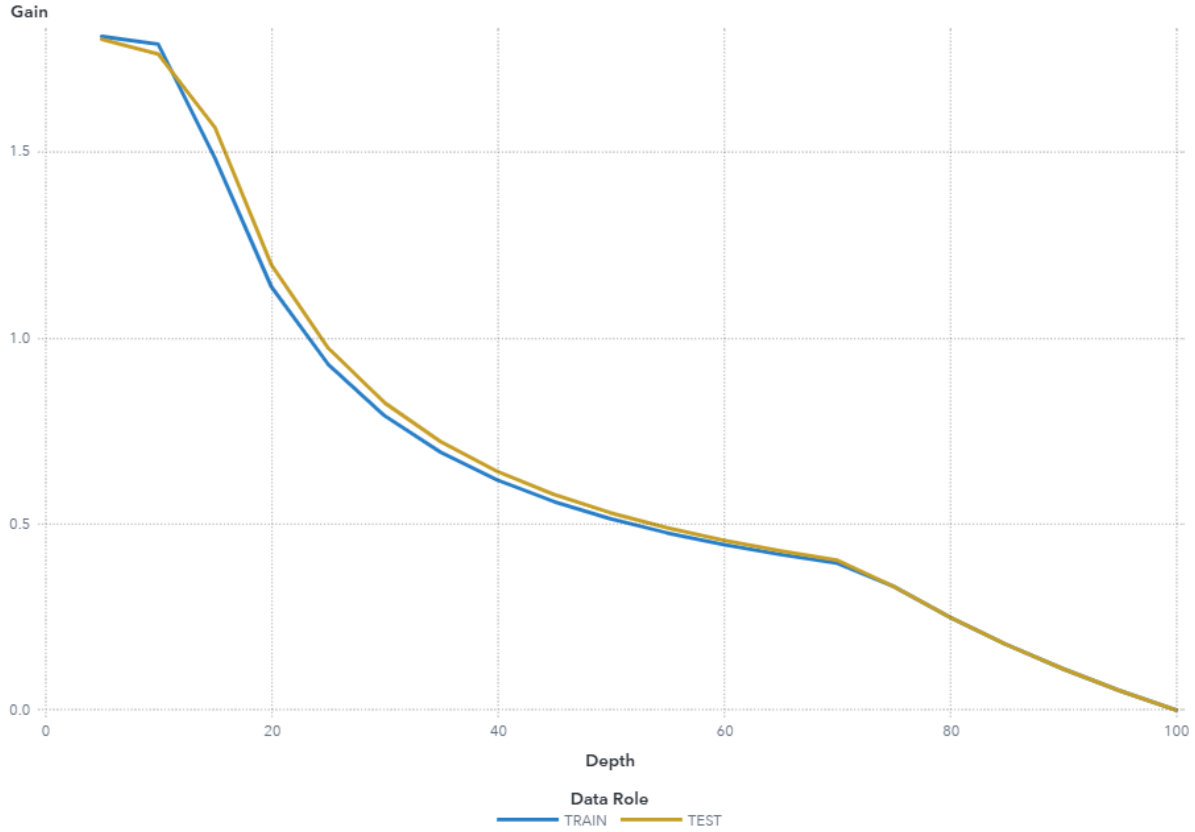


The TRAIN partition has a Lift of 2.81 in the 5% quantile (depth of 5) meaning there are 2.81 times more events in that quantile than expected by random (5% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

The TEST partition has a Lift of 2.8 in the 5% quantile (depth of 5) meaning there are 2.8 times more events in that quantile than expected by random (5% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition.

Lift is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Lift is the ratio of the number of events in that quantile to the number of events that would be there at random, or equivalently, the ratio of the response percentage to the baseline response percentage. With 20 quantiles, it is expected that 5% of the events occur in each quantile. Thus, Lift measures how much more likely it is to observe an event in each quantile than by selecting observations at random.

Gain

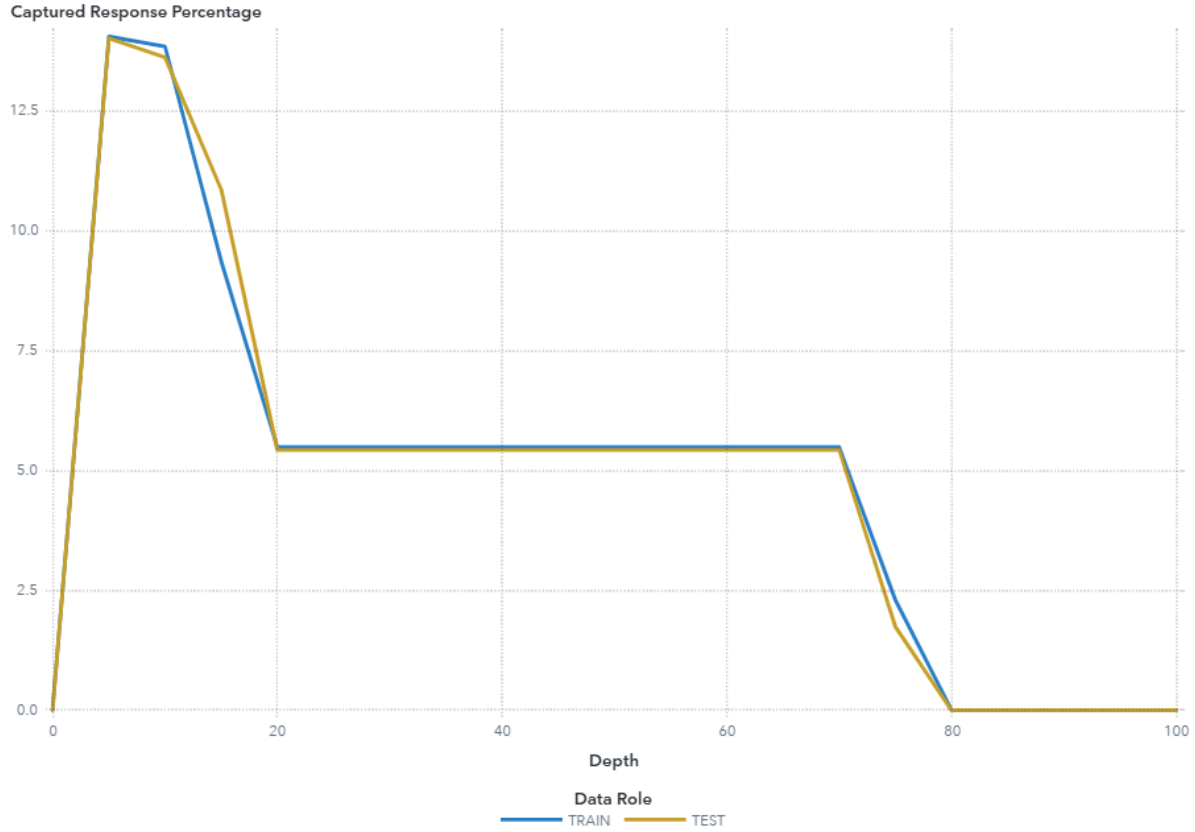


The TRAIN partition has a Gain of 1.8 at the 10% quantile (depth of 10). Because this value is greater than 0, it is better to use your model to identify responders than no model, based on the selected partition. The best possible value of Gain for this partition at depth 10 is 1.81.

The TEST partition has a Gain of 1.8 at the 10% quantile (depth of 10). Because this value is greater than 0, it is better to use your model to identify responders than no model, based on the selected partition. The best possible value of Gain for this partition at depth 10 is 1.8.

Gain is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Gain is a cumulative measure for the quantiles up to an including the current one and is calculated as (number of events in the quantiles) / (number of events expected by random) - 1. With 20 quantiles, it is expected that 5% of the events occur in each quantile. Note that the value of Gain is the same as the value of Cumulative Lift - 1. If the value of Gain is greater than 0, then your model is better at identifying events than using no model.

Captured Response Percentage

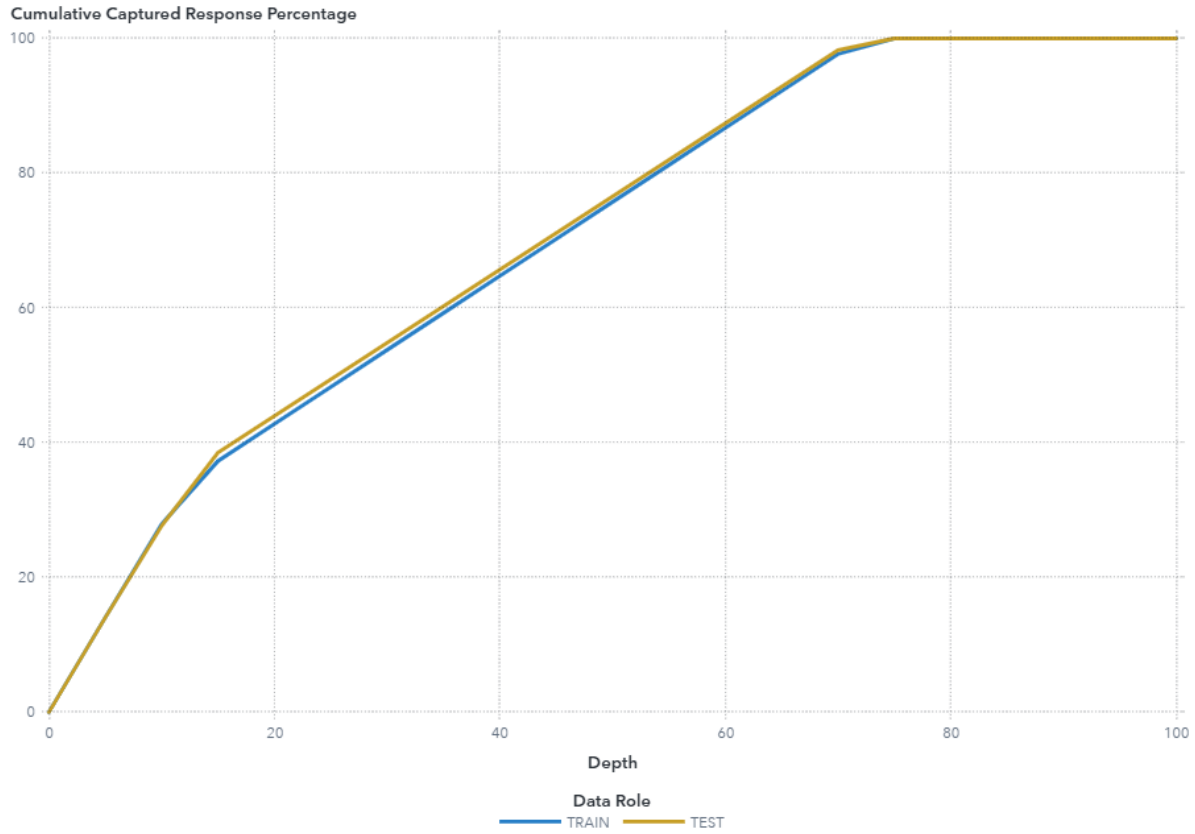


At the 5% quantile (depth of 5), the TRAIN partition has a Captured response percentage of 14.1 (compared to the expected value of 5 for no model). The best possible value of Captured response percentage for this partition at depth 5 is 14.06.

At the 5% quantile (depth of 5), the TEST partition has a Captured response percentage of 14 (compared to the expected value of 5 for no model). The best possible value of Captured response percentage for this partition at depth 5 is 14.02.

Captured response percentage is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Captured response percentage is the percentage of the total number of events that are in that quantile. With no model, it is expected that 5% of the events are in each quantile.

Cumulative Captured Response Percentage

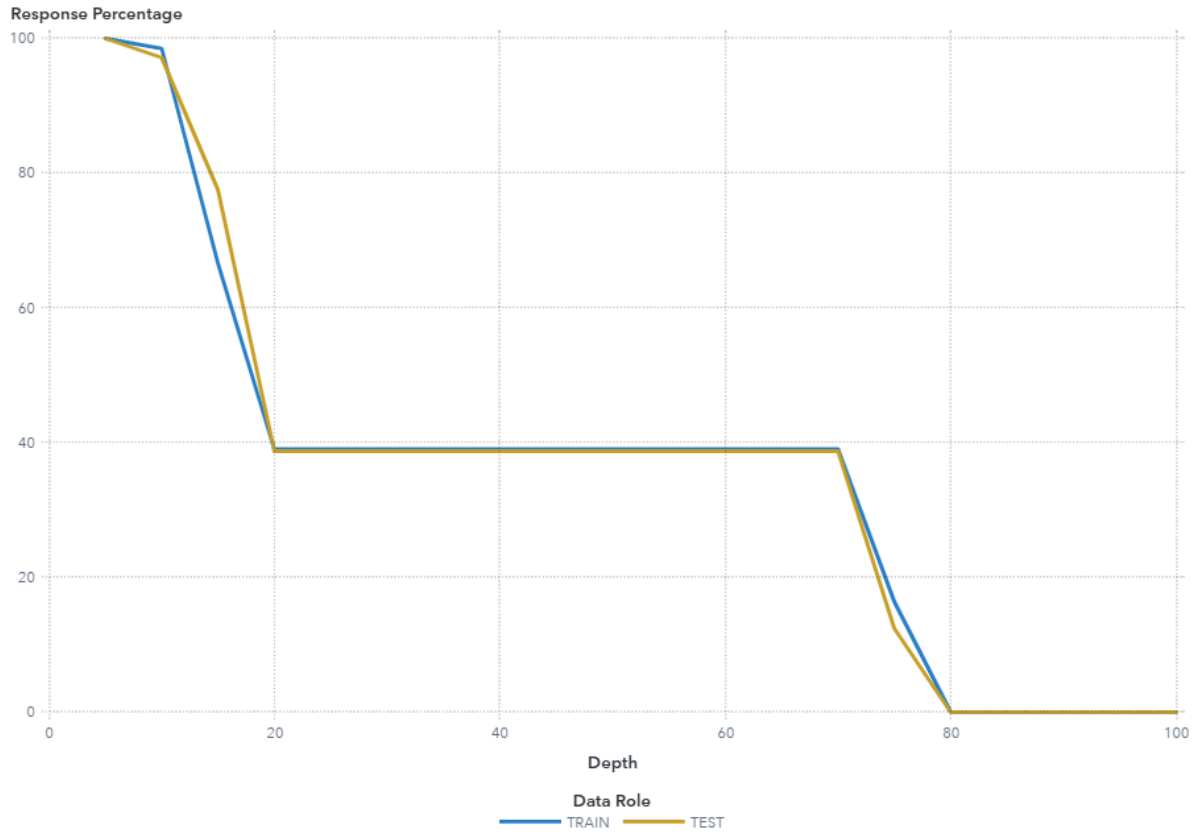


In the top 10% of the data (depth 10), the TRAIN partition has a Cumulative captured response percentage of 27.9 (compared to the expected value of 10 for no model). The best possible value of Cumulative captured response percentage for this partition at depth 10 is 28.12.

In the top 10% of the data (depth 10), the TEST partition has a Cumulative captured response percentage of 27.6 (compared to the expected value of 10 for no model). The best possible value of Cumulative captured response percentage for this partition at depth 10 is 28.04.

Cumulative captured response percentage is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. The cumulative captured response percentage for a particular quantile is the percentage of the total number of events that are in the quantiles up to and including the current quantile. With no model, it is expected that 5% of the events are in each quantile, so the cumulative captured response percentage at depth 10 would be 10%.

Response Percentage

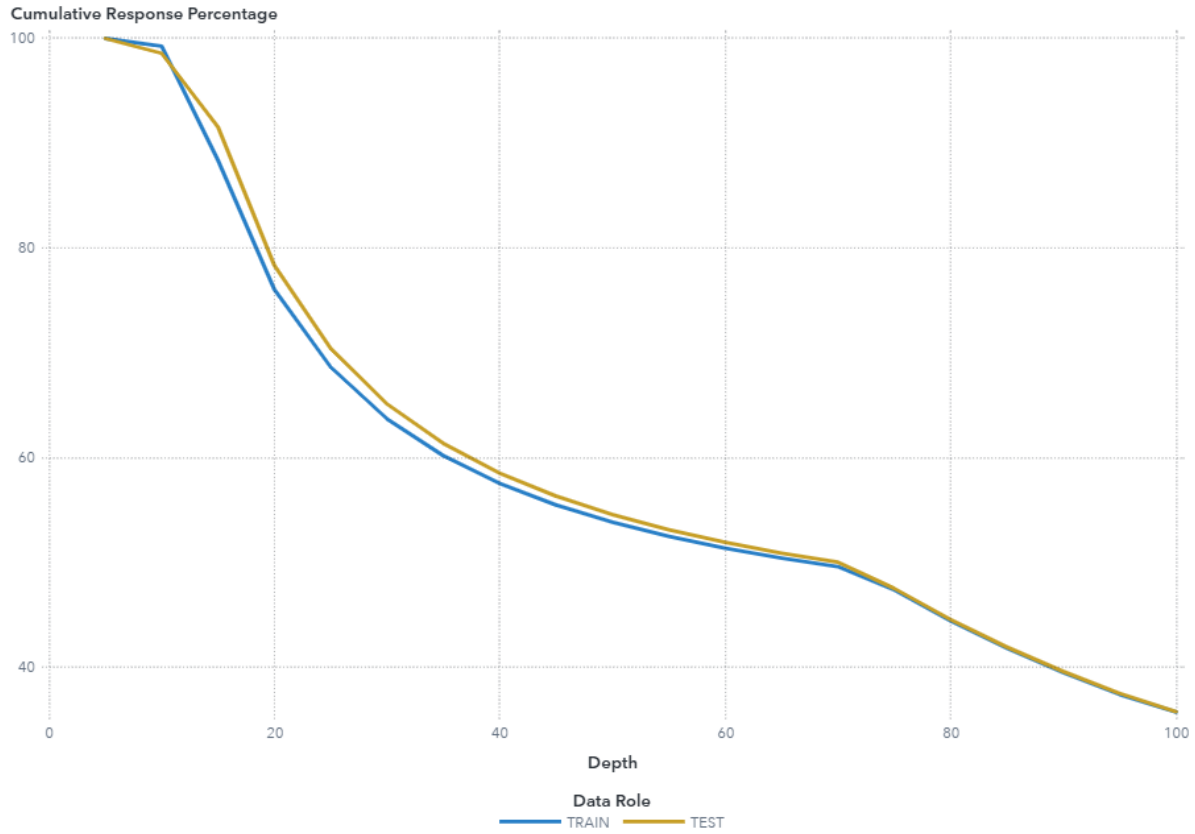


At the 5% quantile (depth of 5), the TRAIN partition has a Response percentage of 100. The best possible value of Response percentage for this partition at depth 5 is 100.

At the 5% quantile (depth of 5), the TEST partition has a Response percentage of 100. The best possible value of Response percentage for this partition at depth 5 is 100.

Response percentage is calculated by sorting each partition in descending order by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Response percentage is the percentage of observations that are events in that quantile. With no model, it is expected that the response percentage is constant across quantiles, 100*overall-event-rate. This is also called the baseline response percentage.

Cumulative Response Percentage

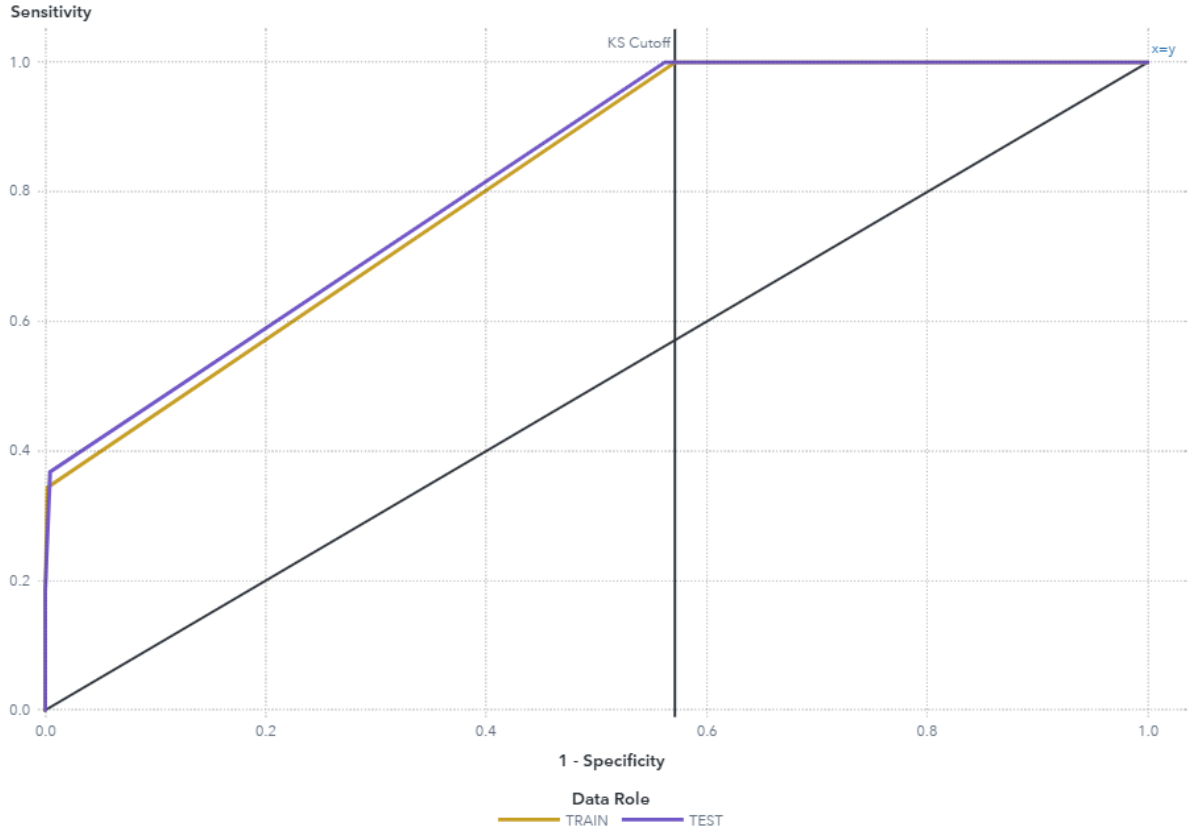


In the top 10% of the data (depth 10), the TRAIN partition has a Cumulative response percentage of 99.3. The best possible value of Cumulative response percentage for this partition at depth 10 is 100.

In the top 10% of the data (depth 10), the TEST partition has a Cumulative response percentage of 98.6. The best possible value of Cumulative response percentage for this partition at depth 10 is 100.

Cumulative response percentage is calculated by sorting in descending order each partition of the data by the predicted probability of the target event P_acci_severity1, which represents the predicted probability of the event "1" for the target acci_severity. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. The cumulative response percentage for a particular quantile is the percentage of observations that are events in the quantiles up to and including the current quantile. With no model, it is expected that the response percentage is constant across quantiles, 100*overall-event-rate. This is also called the baseline response percentage.

ROC



The ROC curve is a plot of sensitivity (the true positive rate) against 1-specificity (the false positive rate), which are both measures of classification based on the confusion matrix. These measures are calculated at various cutoff values. To help identify the best cutoff to use when scoring your data, the KS Cutoff reference line is drawn at the value of 1-specificity where the greatest difference between sensitivity and 1-specificity is observed for the TRAIN partition. The KS Cutoff line is drawn at the cutoff value 0.01, where the 1-specificity value is 0.572 and the sensitivity value is 1.

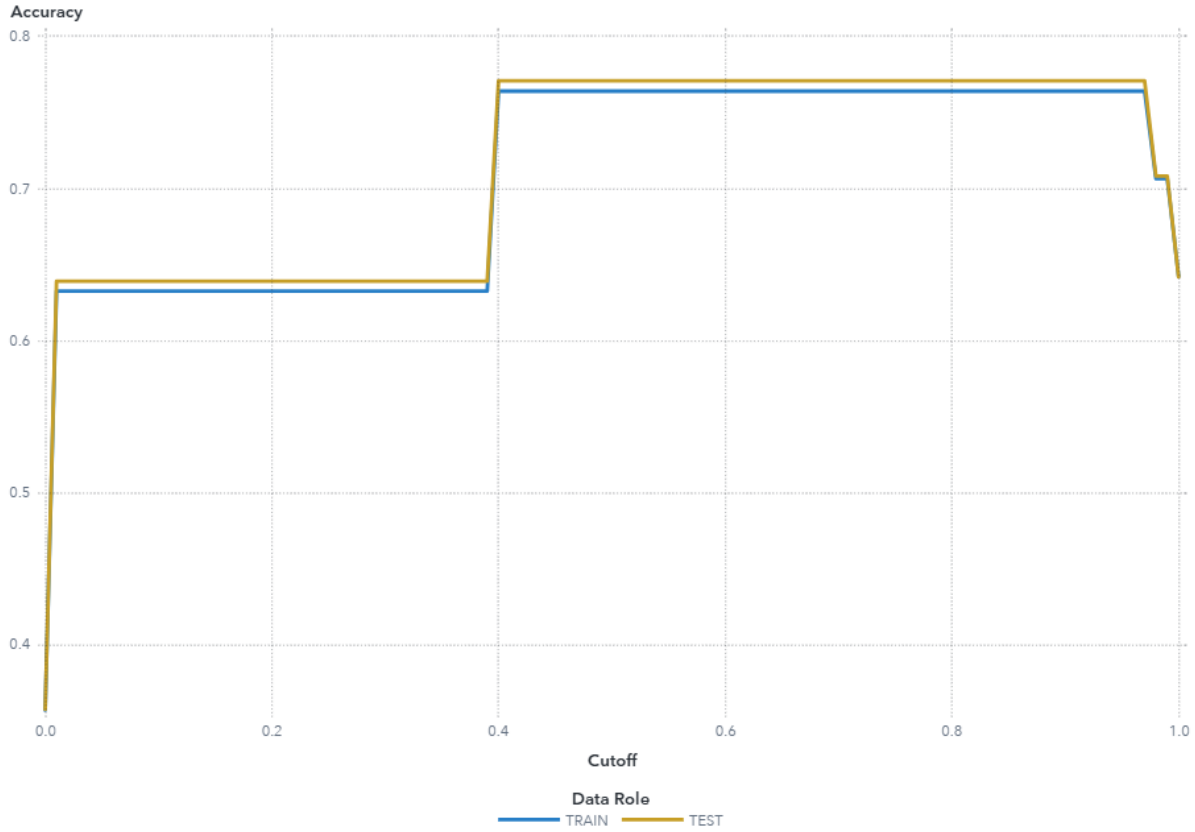
Cutoff values range from 0 to 1, inclusive, in increments of 0.01. At each cutoff value, the predicted target classification is determined by whether P_acci_severity1, which is the predicted probability of the event "1" for the target acci_severity, is greater than or equal to the cutoff value. When P_acci_severity1 is greater than or equal to the cutoff value, then the predicted classification is the event, otherwise it is a non-event.

The confusion matrix for each cutoff value contains four cells that display the true positives for events that are correctly classified (TP), false positives for non-events that are classified as events (FP), false negatives for events that are classified as non-events (FN), and true negatives for non-events that are classified as non-events (TN). True negatives include non-event classifications that specify a different non-event. Sensitivity is calculated as TP / (TP + FN). Specificity, the true negative rate, is calculated as TN / (TN + FP), so 1-specificity is FP / (TN + FP). The values of

sensitivity and 1-specificity are plotted at each cutoff value.

A ROC curve that rapidly approaches the upper-left corner of the graph, where the difference between sensitivity and 1-specificity is the greatest, indicates a more accurate model. A diagonal line where sensitivity = 1-specificity indicates a random model.

Accuracy

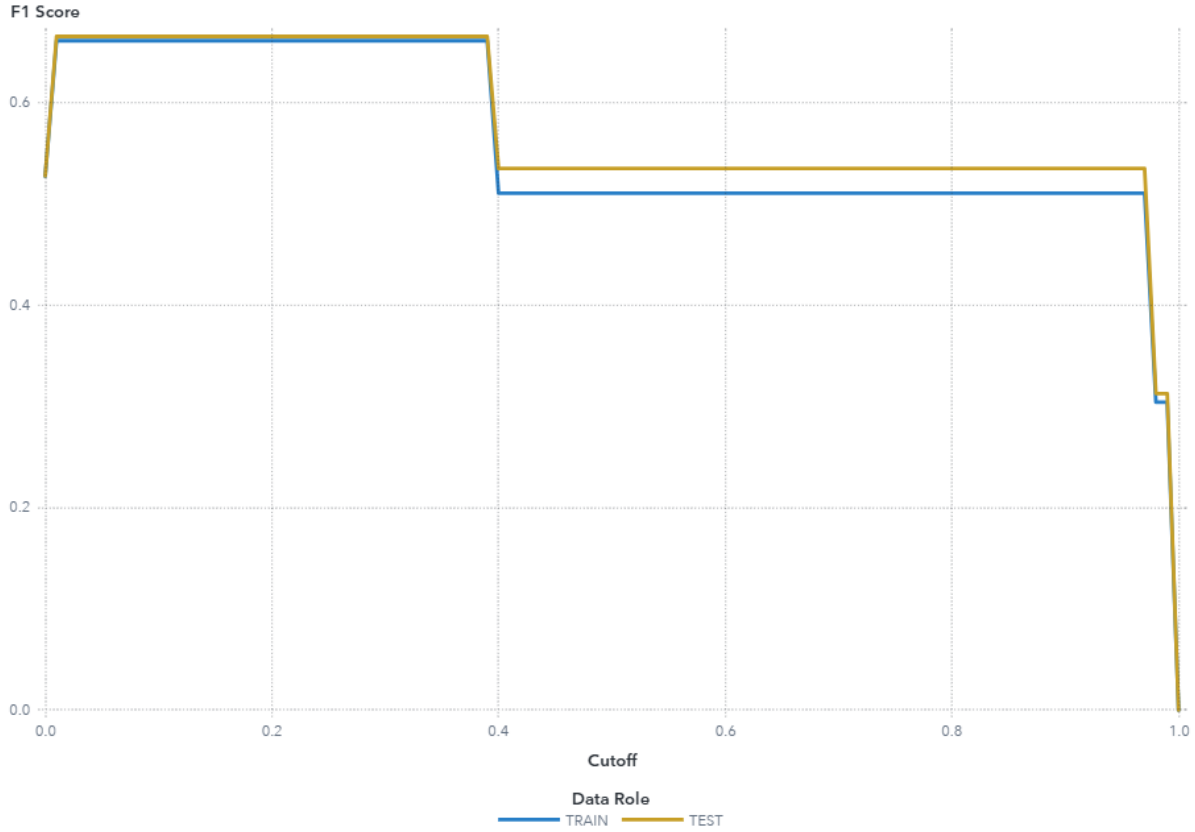


For this model, the accuracy in the TEST partition at the cutoff of 0.5 is 0.771.

For this model, the accuracy in the TRAIN partition at the cutoff of 0.5 is 0.764.

Accuracy is the proportion of observations that are correctly classified as either an event or non-event, calculated at various cutoff values. Cutoff values range from 0 to 1, inclusive, in increments of 0.01. At each cutoff value, the predicted target classification is determined by whether P_acci_severity1, which is the predicted probability of the event "1" for the target acci_severity, is greater than or equal to the cutoff value. When P_acci_severity1 is greater than or equal to the cutoff value, then the predicted classification is the event, otherwise it is a non-event. When the predicted classification and the actual classification are both events (true positives) or both non-events (true negatives), the observation is correctly classified. If the predicted classification and actual classification disagree, then the observation is incorrectly classified. Accuracy is calculated as (true positives + true negatives) / (total observations).

F1 Score



For this model, the F1 score in the TEST partition at the cutoff of 0.5 is 0.535.

For this model, the F1 score in the TRAIN partition at the cutoff of 0.5 is 0.51.

The F1 score combines the measures of precision and recall (or sensitivity), which are measures of classification based on the confusion matrix that are calculated at various cutoff values. Cutoff values range from 0 to 1, inclusive, in increments of 0.01. At each cutoff value, the predicted target classification is determined by whether P_acci_severity1, which is the predicted probability of the event "1" for the target acci_severity, is greater than or equal to the cutoff value. When P_acci_severity1 is greater than or equal to the cutoff value, then the predicted classification is the event, otherwise it is a non-event.

The confusion matrix for each cutoff value contains four cells that display the true positives for events that are correctly classified (TP), false positives for non-events that are classified as events (FP), false negatives for events that are classified as non-events (FN), and true negatives for non-events that are classified as non-events (TN). True negatives include non-event classifications that specify a different non-event.

Precision is calculated as TP / (TP + FP), and recall (or sensitivity) is calculated as

TP / (TP + FN). The F1 score is calculated as 2*Precision*Recall / (Precision + Recall), which is the harmonic mean of Precision and Recall. Larger F1 scores indicate a more accurate model.

Fit Statistics

Target Name	Data Role	Partition Indicator	Formatted Partition
acci_severity	TEST	2	2
acci_severity	TRAIN	1	1

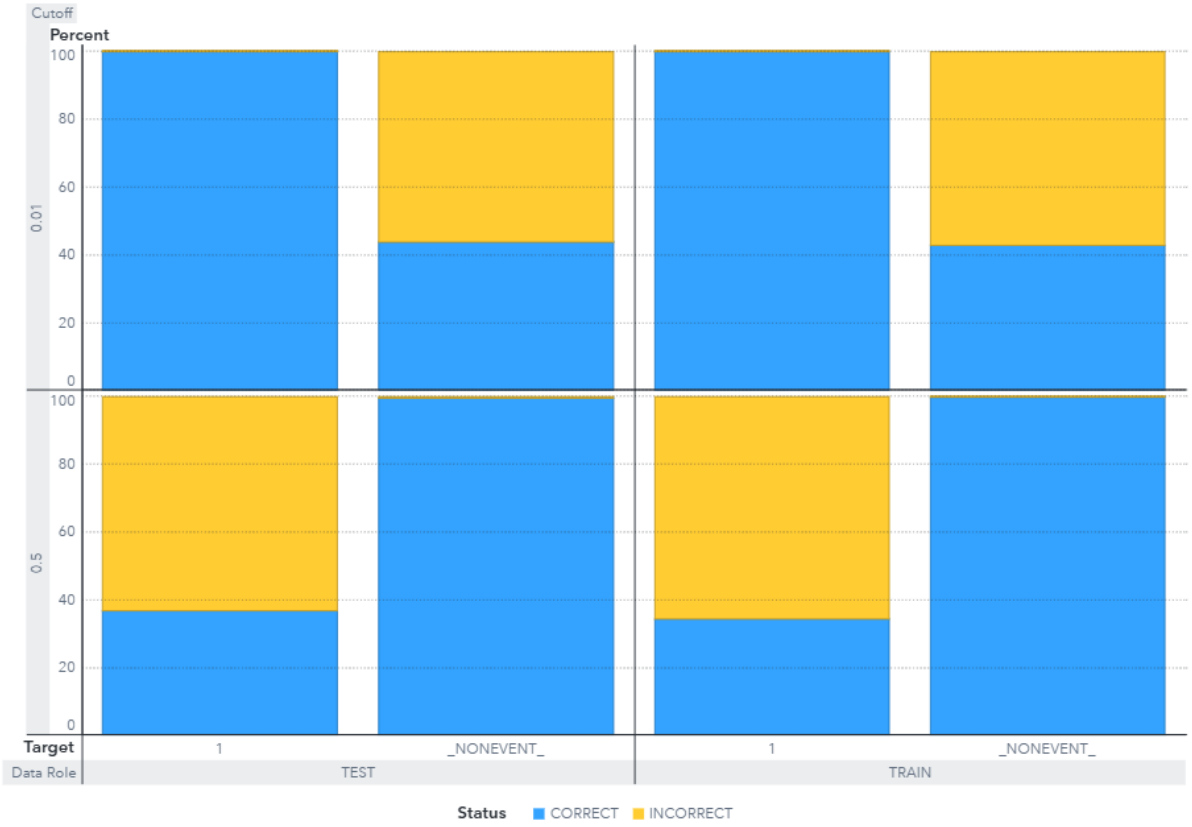
Number of Observations	Average Squared Error	Divisor for ASE	Root Average Squared Error
1,056	0.1768	1,056	0.4205
2,467	0.1736	2,467	0.4167

Misclassification Rate	Multi-Class Log Loss	KS (Youden)	Area Under ROC
0.4782	0.8457	0.4381	0.8205
0.4682	0.8342	0.4284	0.8116

Gini Coefficient	Gamma	Tau	KS Cutoff
0.6411	0.9913	0.2950	0.0100
0.6232	0.9960	0.2864	0.0100

KS at Default Cutoff	Misclassification Rate at KS Cutoff (Event)	Misclassification Rate (Event)
0.3633	0.3608	0.2292
0.3416	0.3672	0.2359

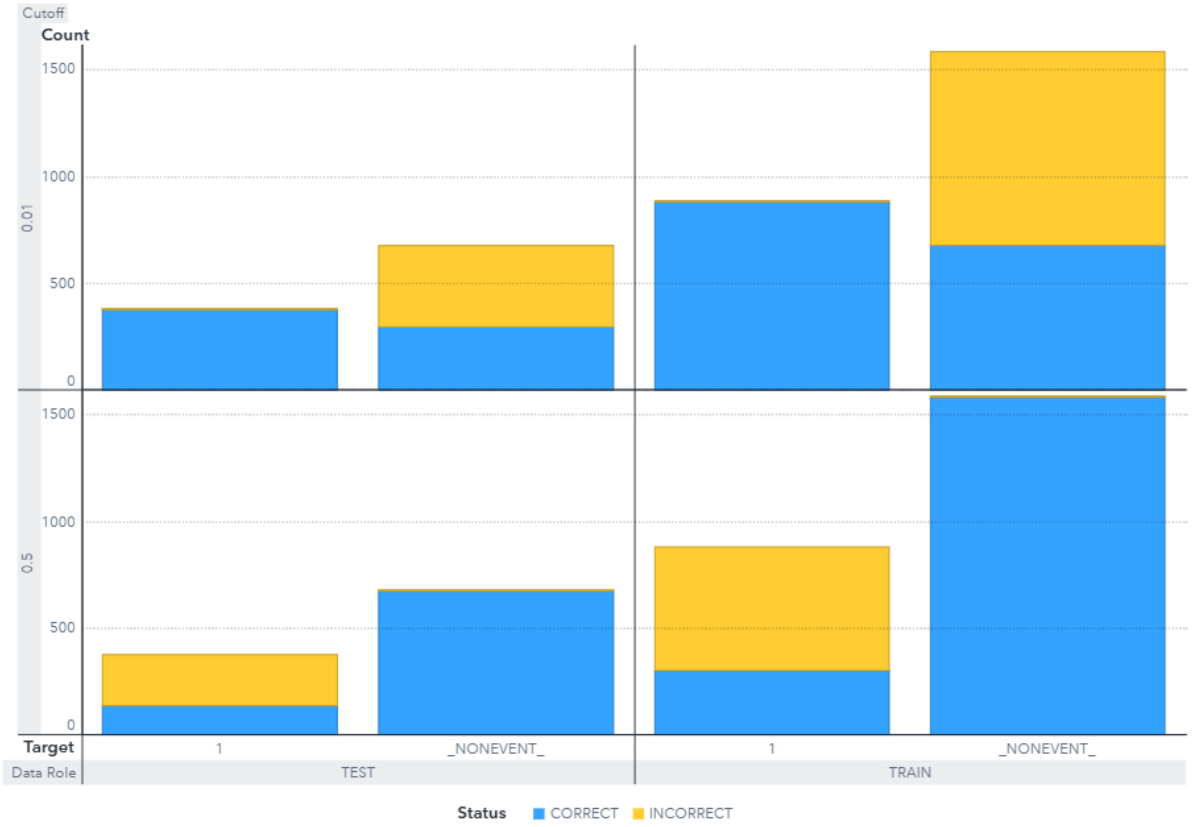
Percentage Plot



The Event Classification report is a visual representation of the confusion matrix at various cutoff values for each partition. The classification cutoffs used in the plot are the default (0.5) and these KS cutoff values for existing partitions: 0.01 (TRAIN), 0.01 (TEST).

For this data, for the bar corresponding to the event level of acci_severity, " 1", the segment of the bar colored as "CORRECT" corresponds to true positives.

Count Plot



The Event Classification report is a visual representation of the confusion matrix at various cutoff values for each partition. The classification cutoffs used in the plot are the default (0.5) and these KS cutoff values for existing partitions: 0.01 (TRAIN), 0.01 (TEST).

For this data, for the bar corresponding to the event level of acci_severity, " 1", the segment of the bar colored as "CORRECT" corresponds to true positives.

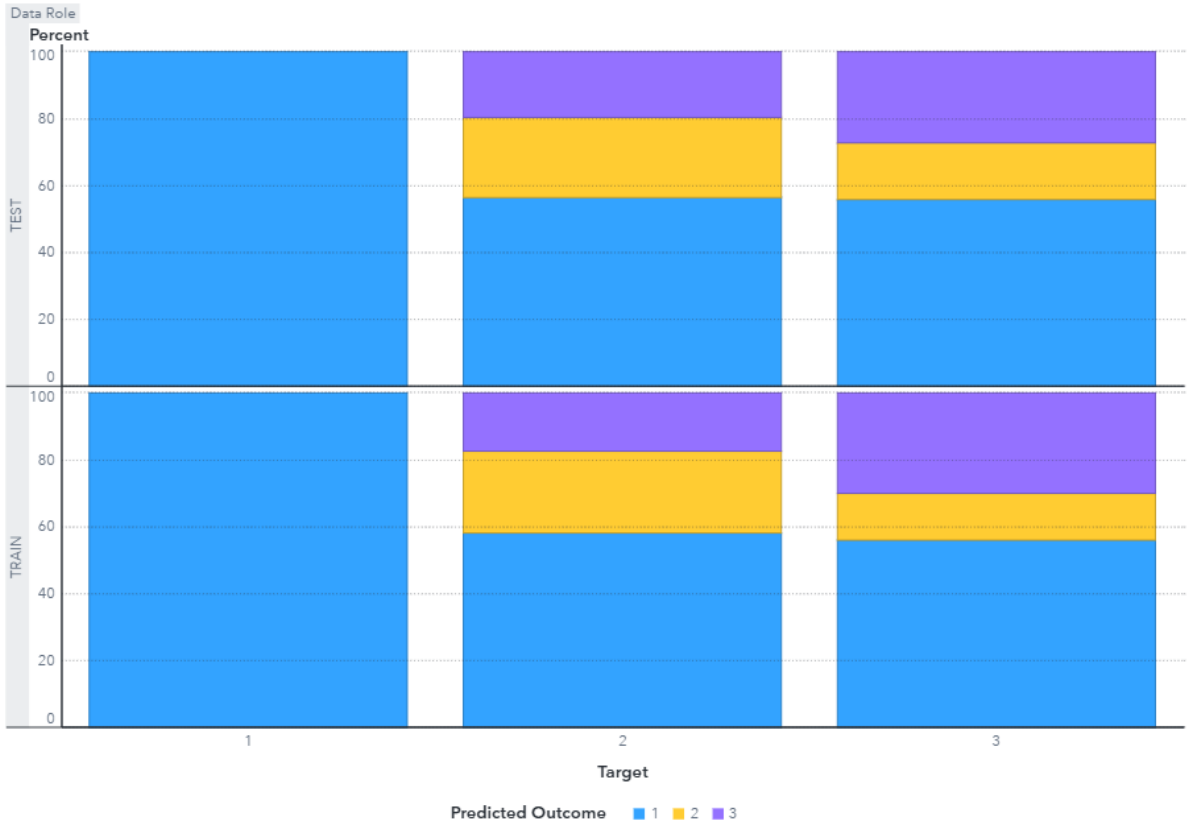
Table

Cutoff	Cutoff Source	Target Name	Response
0.0100	KS	acci_severity	CORRECT
0.0100	KS	acci_severity	INCORRECT
0.0100	KS	acci_severity	CORRECT
0.0100	KS	acci_severity	INCORRECT
0.5000	Default	acci_severity	CORRECT
0.5000	Default	acci_severity	INCORRECT
0.5000	Default	acci_severity	CORRECT
0.5000	Default	acci_severity	INCORRECT

Event	Value	Training Frequency	Validation Frequency
1	True Positive	882	
1	False Negative	0	
NONEVENT	True Negative	679	
NONEVENT	False Positive	906	
1	True Positive	303	
1	False Negative	579	
NONEVENT	True Negative	1,582	
NONEVENT	False Positive	3	

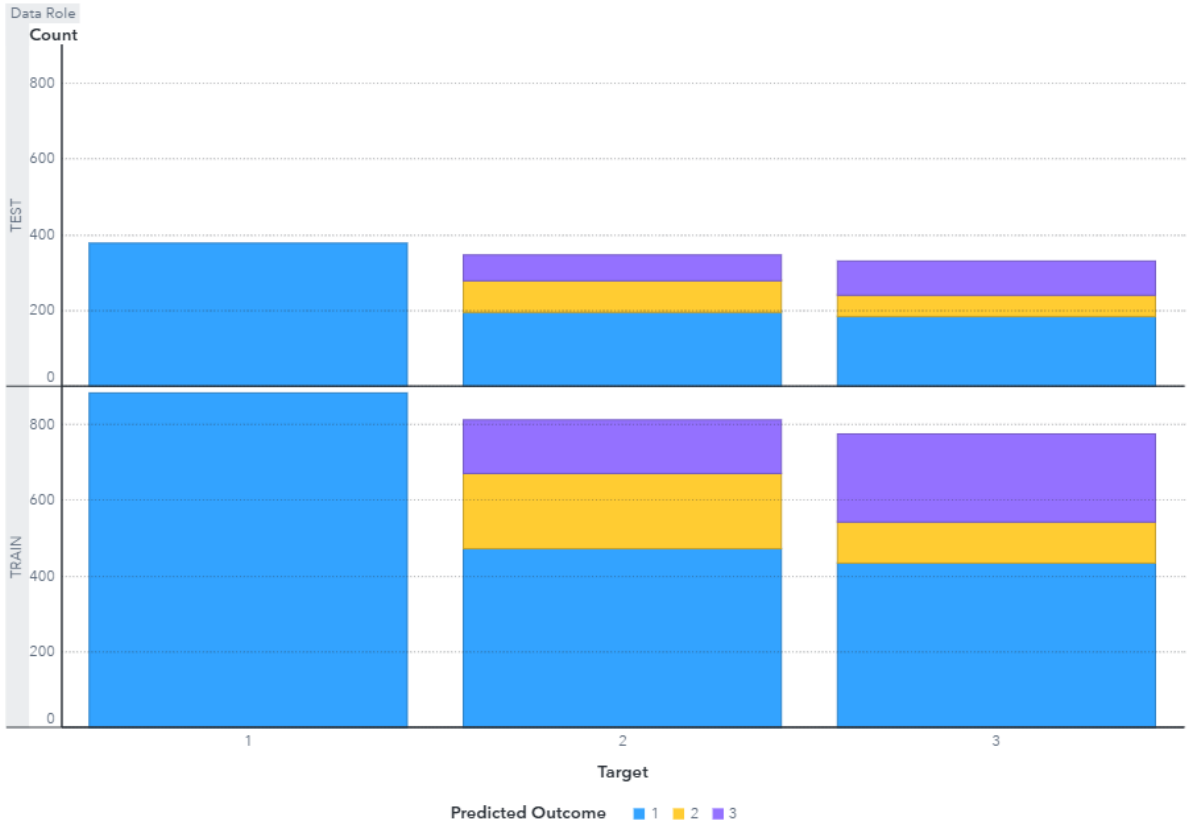
Test Frequency	Training Percentage	Validation Percentage	Test Percentage
378	100		100
0	0		0
297	42.8391		43.8053
381	57.1609		56.1947
139	34.3537		36.7725
239	65.6463		63.2275
675	99.8107		99.5575
3	0.1893		0.4425

Percentage Plot



The Nominal Classification report displays either the percentage of or the number of observations predicting each target level. The plot is segmented by target level and partition level. The target level with the greatest predicted probability is the predicted outcome. A greater number of observations where the target and predicted outcome are the same indicates a better model.

Count Plot



The Nominal Classification report displays either the percentage of or the number of observations predicting each target level. The plot is segmented by target level and partition level. The target level with the greatest predicted probability is the predicted outcome. A greater number of observations where the target and predicted outcome are the same indicates a better model.

Table

Target Name	Data Role	Target	Unformatted Target
acci_severity	TRAIN	1	1
acci_severity	TRAIN	2	2
acci_severity	TRAIN	2	2
acci_severity	TRAIN	2	2
acci_severity	TRAIN	3	3
acci_severity	TRAIN	3	3
acci_severity	TRAIN	3	3
acci_severity	TEST	1	1
acci_severity	TEST	2	2
acci_severity	TEST	2	2
acci_severity	TEST	2	2
acci_severity	TEST	3	3
acci_severity	TEST	3	3
acci_severity	TEST	3	3

Predicted Outcome	Count	Percent	Status
1	882	100	CORRECT
1	472	58.1998	INCORRECT
2	198	24.4143	CORRECT
3	141	17.3859	INCORRECT
1	434	56.0724	INCORRECT
2	108	13.9535	INCORRECT
3	232	29.9742	CORRECT
1	378	100	CORRECT
1	196	56.4841	INCORRECT
2	83	23.9193	CORRECT
3	68	19.5965	INCORRECT
1	185	55.8912	INCORRECT
2	56	16.9184	INCORRECT

Predicted Outcome	Count	Percent	Status
3	90	27.1903	CORRECT

Properties

Property Name	Property Value
alpha	0.2000
atAppendLookup	false
atCreateHistory	false
atHistoryLibUri	
atHistoryTblName	
atLeaveAutotuneOn	false
atLookupTableUri	
atMaxBayes	100
atMaxEval	50
atMaxIter	5
atMaxTime	60
atObjectiveInt	ASE
atObjectiveNom	KS
atPopSize	10
atSampleSize	50
atSearchMethod	GA
atTrainProp	0.7000
atUpdateProperties	false
atUseLookup	false
atValidFold	5
atValidMethod	PARTITION
atValidProp	0.3000
atgrowcrit	true
atgrowcritValsi	VARIANCE FTEST CHAID
atgrowcritValsn	ENTROPY CHAID IGR GINI CHISQUARE
atleafSize	false

Property Name	Property Value
atleafSizeInit	5
atleafSizeLB	1
atleafSizeUB	100
atmaxdepth	true
atmaxdepthInit	10
atmaxdepthLB	1
atmaxdepthUB	19
atnumbin	true
atnumbinInit	50
atnumbinLB	20
atnumbinUB	200
autotune_enabled	false
binaryProbCutoff	0.5000
bonferroni	false
ccAlpha	0
codeLocation	mlearning
confidence	0.2500
criterionMethod	IGR
cvccFolds	10
dataMiningVersion	V2024.03
editedInteractively	false
embeddedBarChart	true
exactPctlLift	true
explainFidelity	false
explainInfo	false
fullDatasetReconstitution	false
hLeafSize	5
iCriterionMethod	VARIANCE
icePlots	false
inodeColor	AVERAGE

Property Name	Property Value
intBinMethod	QUANTILE
intervalBins	50
maxBranch	2
maxCategories	128
maxDepth	10
maxNumShapVars	20
minUseinsearch	1
missingValue	USEINSEARCH
nBins	50
nPLeaves	1
nodeColor	PROBEVENT
pdNumImportantInputs	5
pdObsSamples	1,000
pdPlots	false
performKernelShap	false
performLime	false
performVI	false
pruningMethod	COSTCOMPLEXITY
rapidGrowth	false
reportingOnly	false
seRule	false
seed	12,345
seedId	12,345
selMethod	AUTOMATIC
specifyRows	RANDOM
templateRevision	4
train	true
truncateLl	5
truncateUl	95

Property Name	Property Value
useVarOnce	false
userProbCutoff	false

Output

The SAS System

The TREESPLIT Procedure

Model Information	
Split Criterion	IGR
Pruning Method	Cost Complexity
Max Branches per Node	2
Max Tree Depth	10
Tree Depth Before Pruning	10
Tree Depth After Pruning	10
Number of Leaves Before Pruning	30
Number of Leaves After Pruning	16

	Training	Test	Total
Number of Observations Read	2467	1056	3523
Number of Observations Used	2467	1056	3523

The SAS System

The TREESPLIT Procedure

10-Fold Cross Validation Assessment of Pruning Parameter

			Misclassification Rate			
N Leaves	Pruning Parameter		Min	Avg	Standard Error	Max
25	2E-11	.	0.3745	0.4494	0.0395	0.5176
21	0.00116	.	0.3786	0.4513	0.0385	0.5176
17	0.00157	.	0.3786	0.4496	0.0395	0.5098
16	0.00222	*	0.3786	0.4479	0.0401	0.5098
14	0.00281	.	0.3868	0.4520	0.0408	0.5137
12	0.00344	.	0.3868	0.4571	0.0432	0.5333
11	0.00421	.	0.4033	0.4632	0.0417	0.5333
10	0.00711	.	0.4074	0.4705	0.0388	0.5373
7	0.0136	.	0.4115	0.5030	0.0538	0.6130
5	0.0197	.	0.4115	0.5355	0.0608	0.6130
3	0.0219	.	0.4115	0.5765	0.0689	0.6453
1	2.2092	.	0.6211	0.6426	0.0198	0.6872

* Selected pruning parameter

The SAS System

The TREESPLIT Procedure

Fit Statistics for Selected Tree		
	Number of Leaves	Misclassification Rate
Training	16	0.4682
Test	16	0.4782

Variable Importance			
	Training		
Variable	Importance	Relative Importance	Count
hour_of_day	85.4301	1.0000	2
junc_detail	45.5964	0.5337	2
loc_auth_ons_distr	43.9653	0.5146	1
longitude	34.5935	0.4049	1
first_road_num	30.3521	0.3553	1
num_of_casu	28.3856	0.3323	1
weath_con	26.5254	0.3105	1
speed_limit	19.9007	0.2329	1
urb_or_rur_area	14.0177	0.1641	1
ped_cross_hum_con	10.8281	0.1267	1
latitude	7.2924	0.0854	1
road_type	7.0448	0.0825	1
carri_haz	3.4867	0.0408	1

The SAS System

The TREESPLIT Procedure

Predicted Probability Variables	
acci_severity	Variable
1	P_acci_severity1
3	P_acci_severity3
2	P_acci_severity2

Predicted Target Variable	
Level Index	Variable
	I_acci_severity

Accident_Fatality_Predict...

"Model Comparison" Results

by: ta01468@surrey.ac.uk

Contents

Model Comparison

Champion	Name	Algorithm Name	KS (Youden)
true	Decision Tree	Decision Tree	0.4381
false	Logistic Regression	Logistic Regression	0.3540

Accuracy	Average Squared Error	Area Under ROC	Cumulative Lift
0.7708	0.1768	0.8205	2.7646
0.6799	0.2056	0.7028	1.9866

Cumulative Captured Response Percentage	Cutoff	Data Role	Depth
27.6455	0.5000	TEST	10
19.8661	0.5000	TEST	10

F1 Score	False Discovery Rate	False Positive Rate	Gain
0.5346	0.0211	0.0044	1.7646
0.4192	0.4020	0.1209	0.9866

Gini Coefficient	ROC Separation	Lift	Misclassification Rate
0.6411	0.3633	2.7249	0.4782
0.4055	0.2018	1.7163	0.5275

Multi-Class Log Loss	Misclassification at Cutoff	Misclassification Rate (Event)	Number of Observations
0.8457	0.4782	0.2292	1,056
1.0232	0.5275	0.3201	1,056

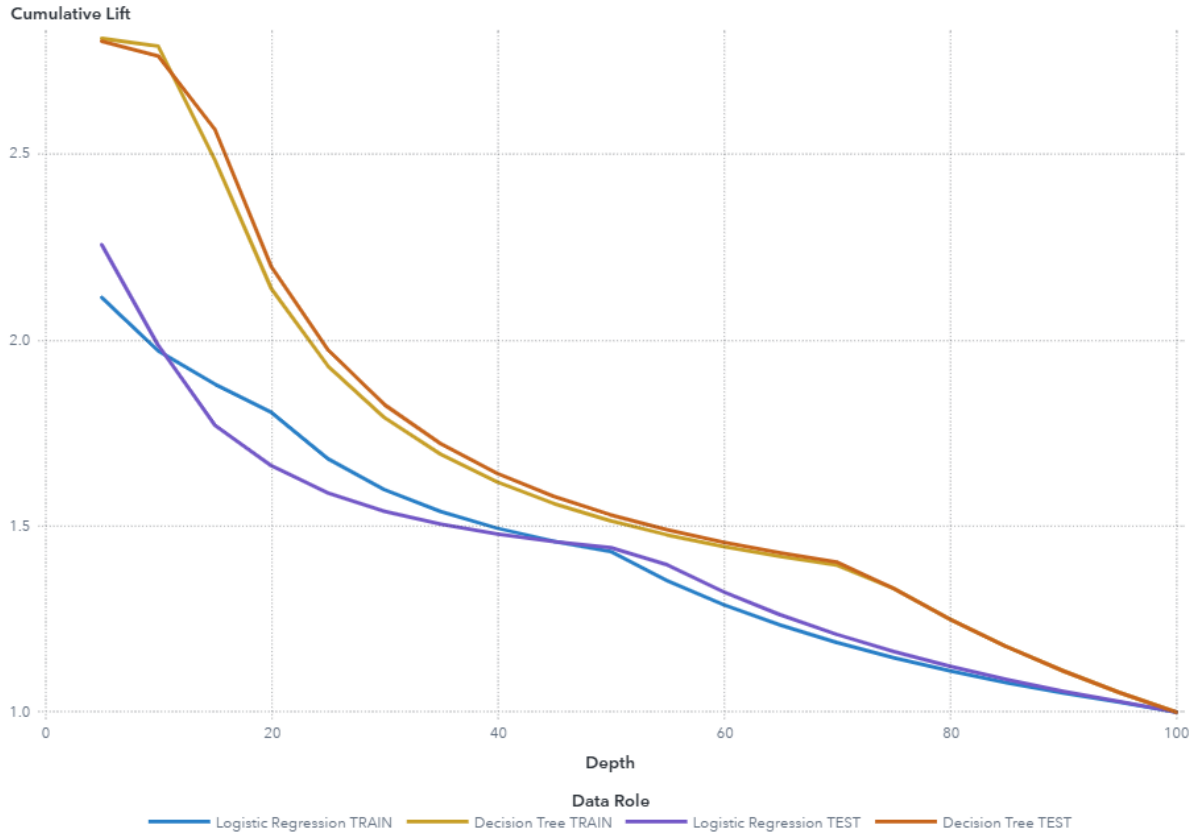
Root Average Squared Error	Captured Response Percentage

Root Average Squared Error	Captured Response Percentage
0.4205	13.6243
0.4534	8.5813

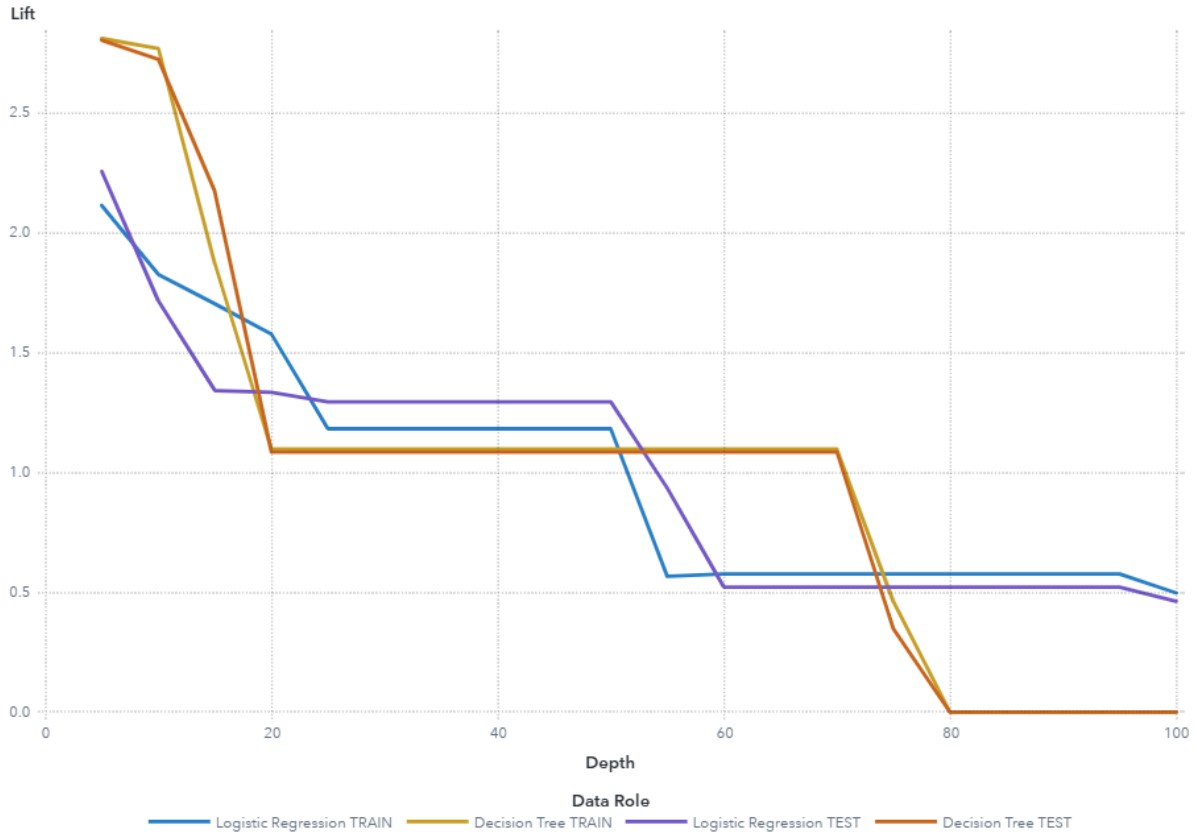
Properties

Property Name	Property Value
selectionCriteriaClass	Kolmogorov-Smirnov statistic (KS)
selectionCriteriaInterval	Average squared error
selectionTable	Test
selectionDepth	10
cutoff	0.50

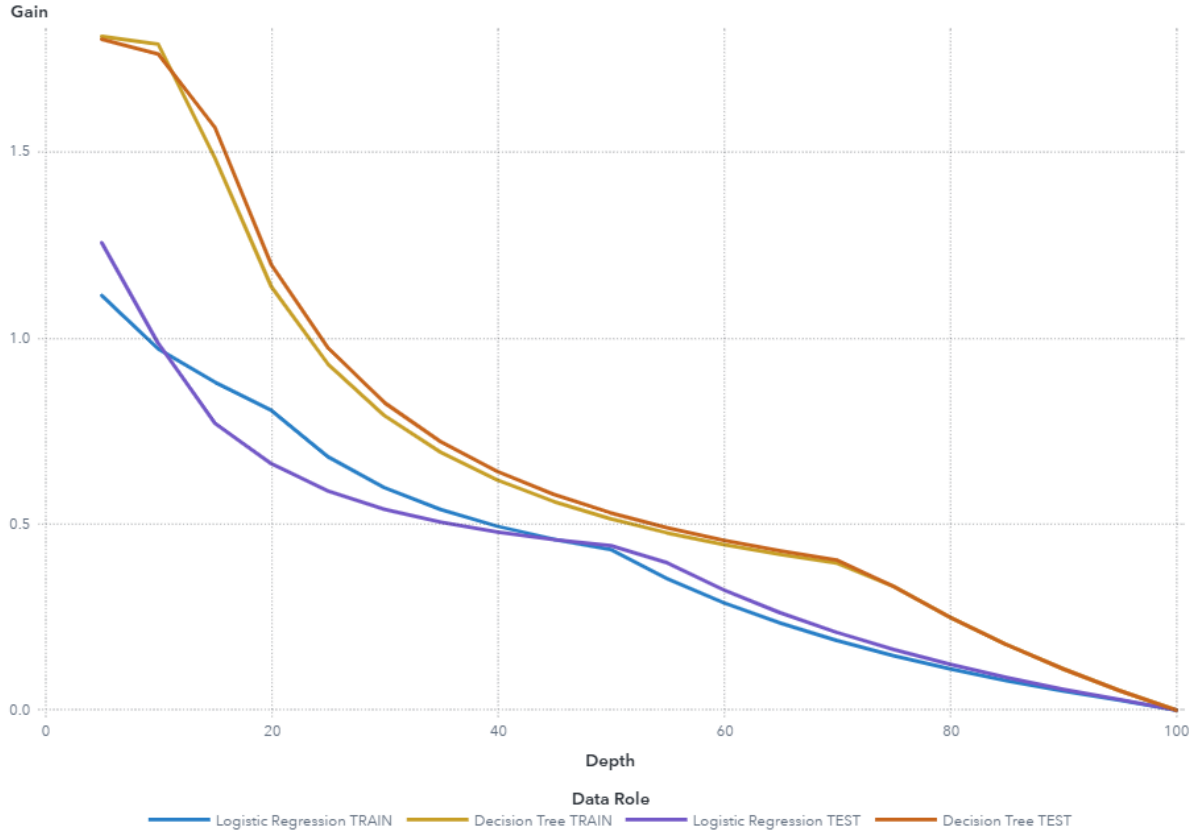
Cumulative Lift



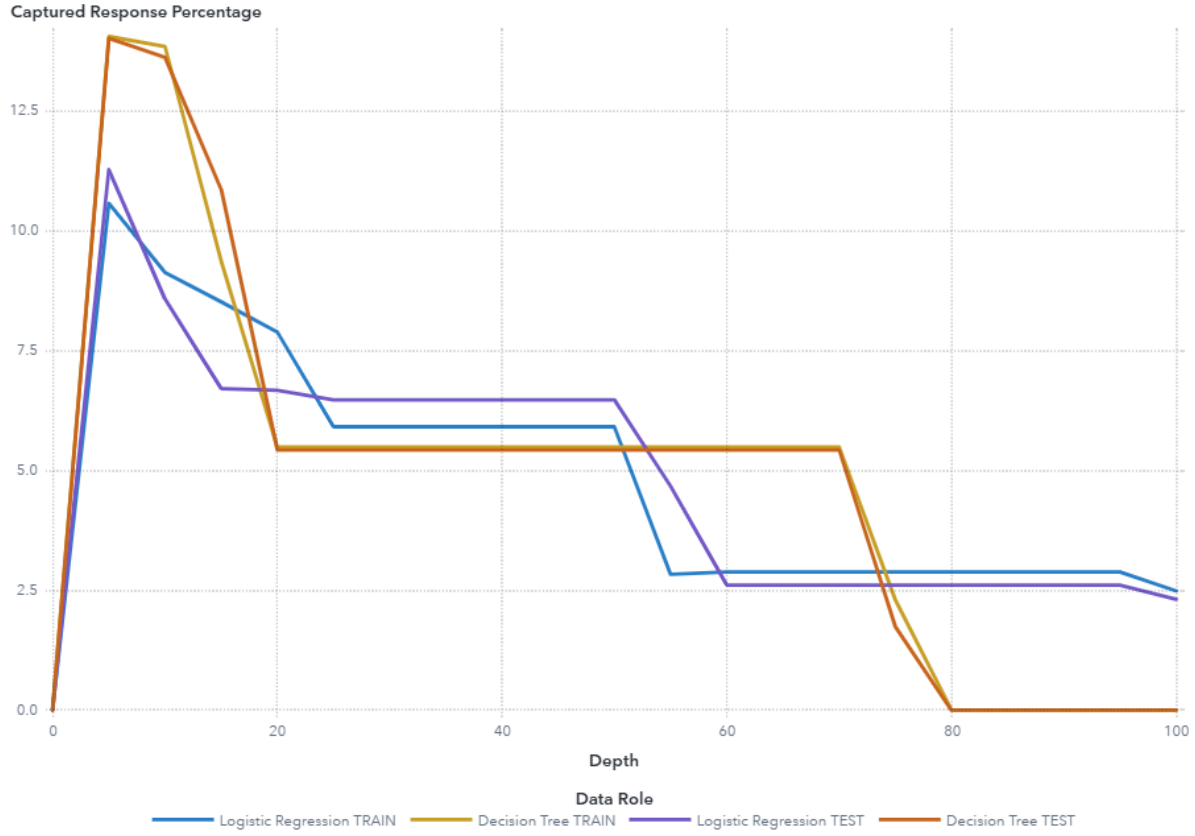
Lift


Lift
2.5
2.0
1.5
1.0
0.5
0.0
0
20
40
60
80
100
Depth
Data Role
Logistic Regression TRAIN
Decision Tree TRAIN
Logistic Regression TEST
Decision Tree TEST

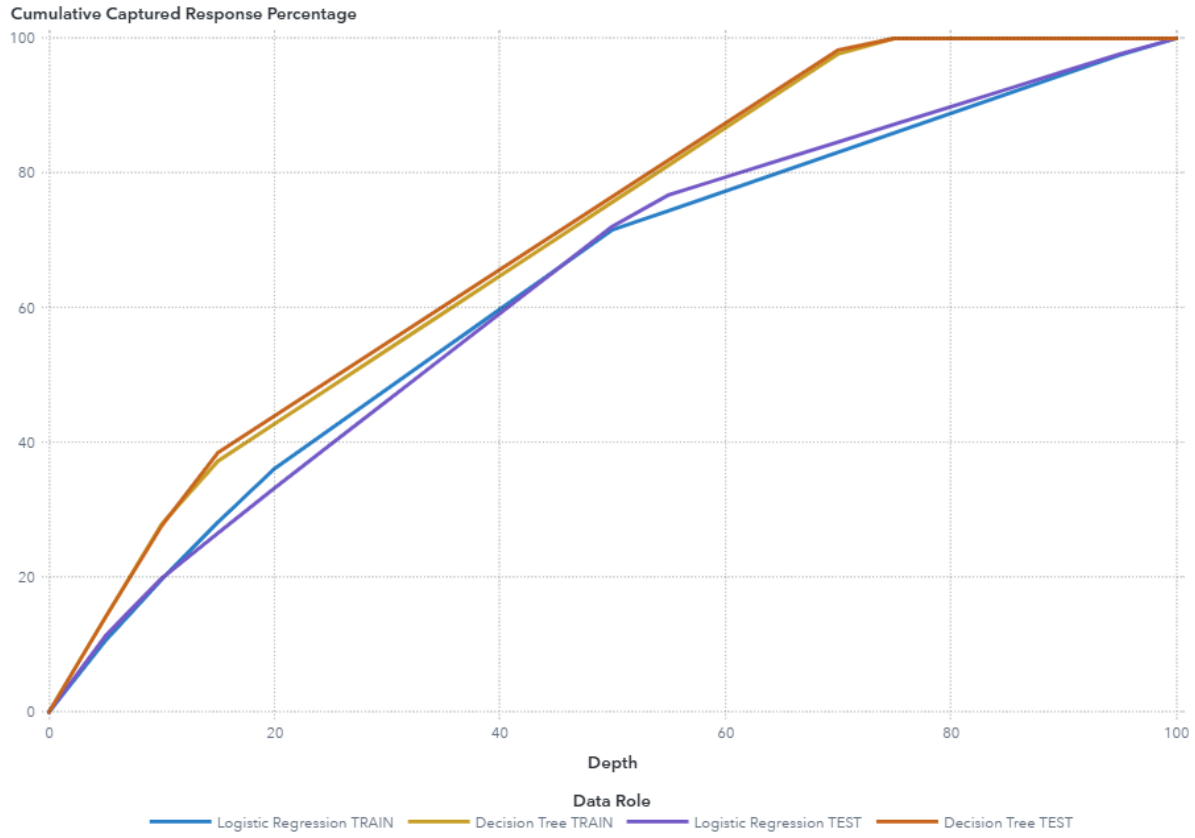
Gain


Gain
1.5
1.0
0.5
0.0
0
20
40
60
80
100
Depth
Data Role
Logistic Regression TRAIN
Decision Tree TRAIN
Logistic Regression TEST
Decision Tree TEST

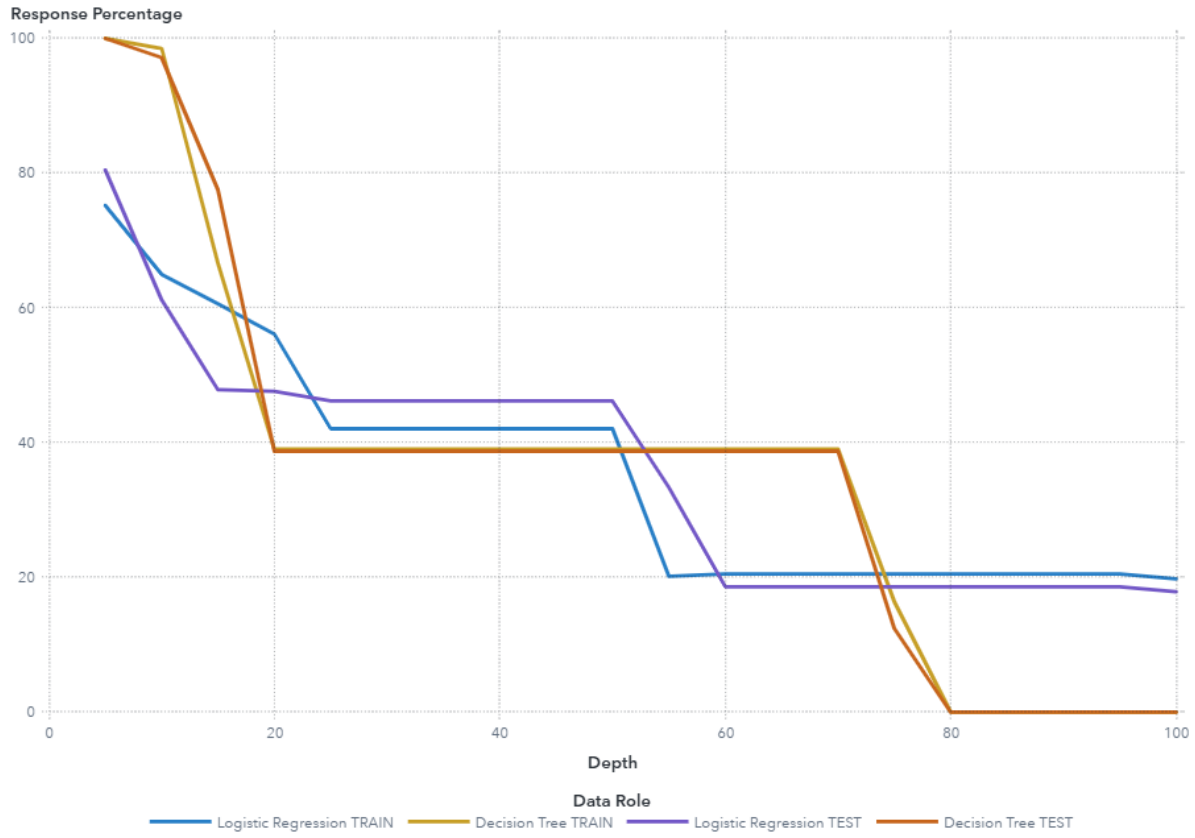
Captured Response Percentage



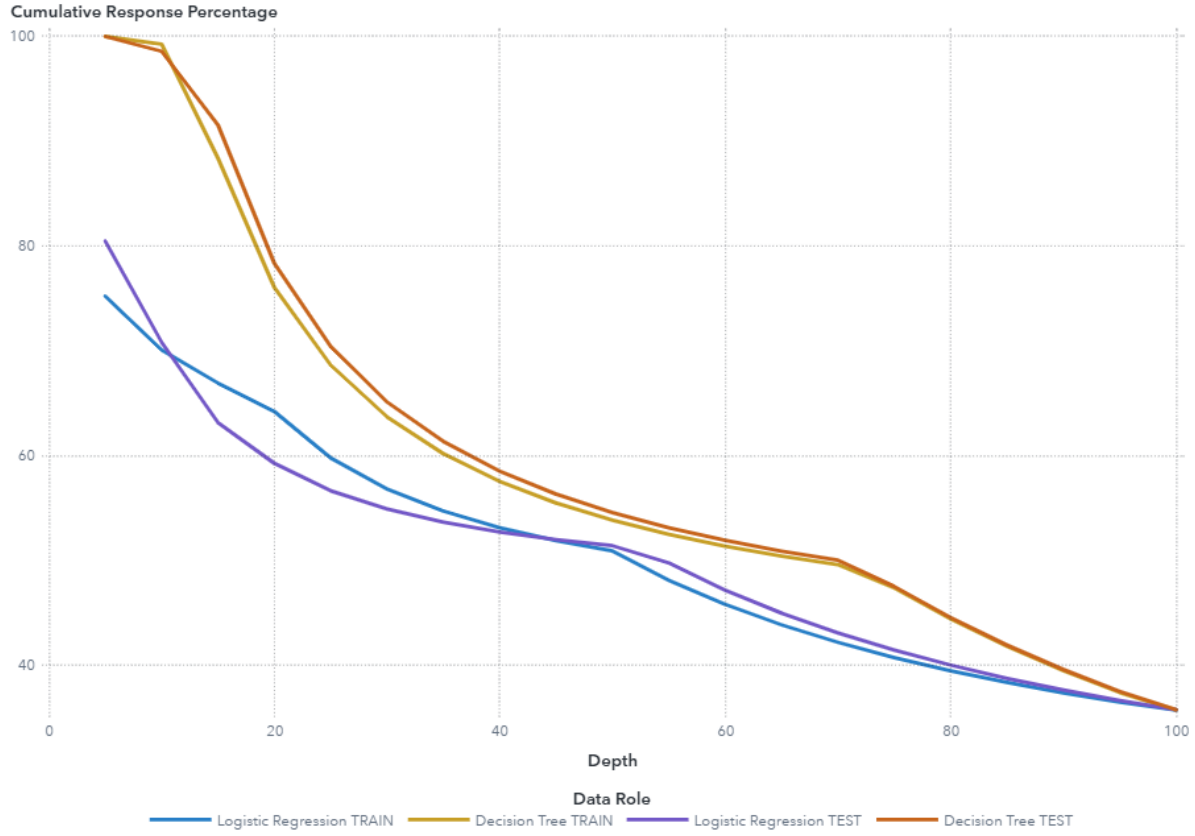
Cumulative Captured Response Percentage



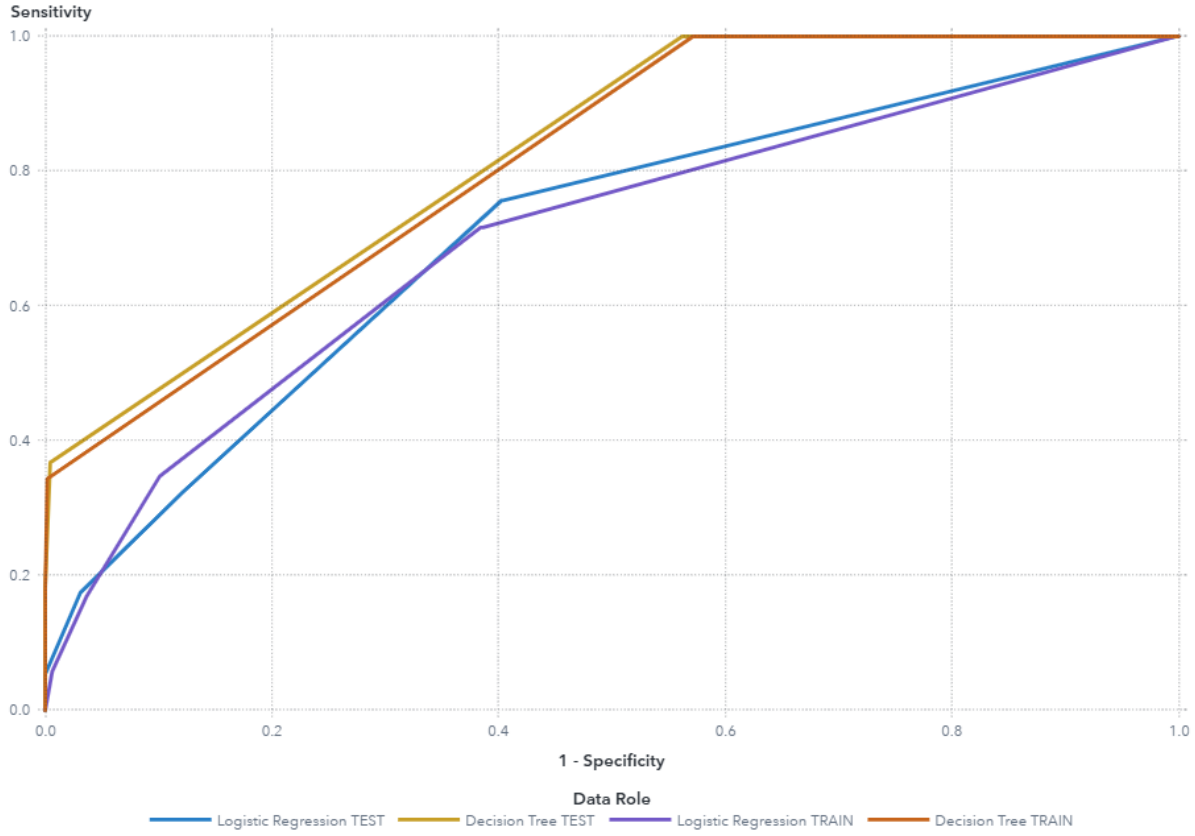
Response Percentage



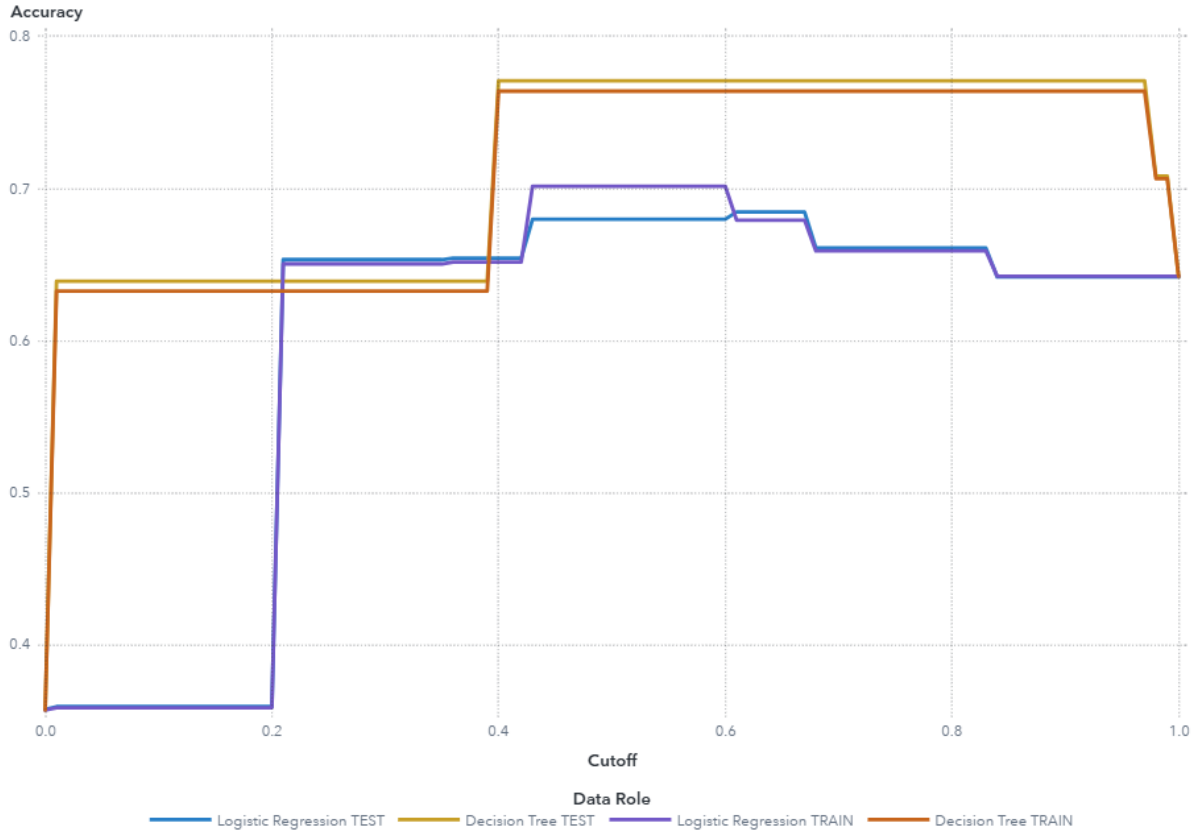
Cumulative Response Percentage



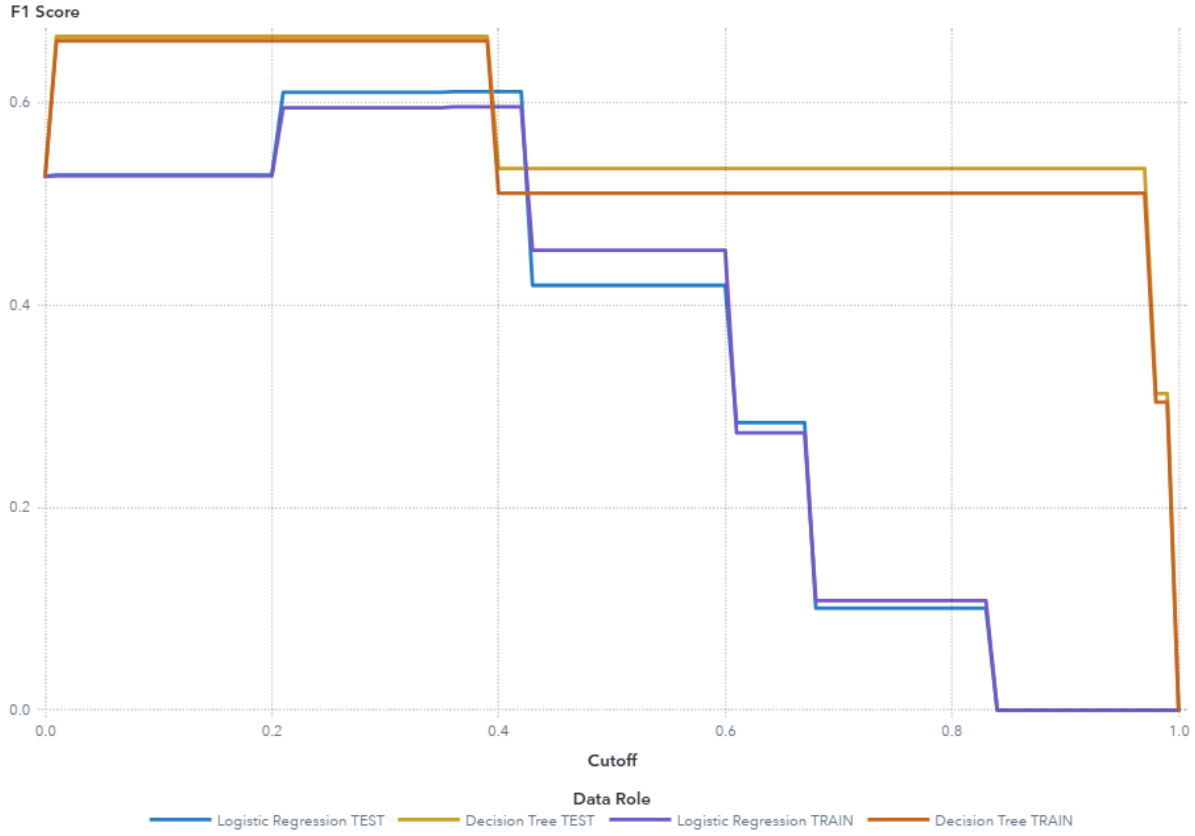
ROC


Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity
Data Role
Logistic Regression TEST
Decision Tree TEST
Logistic Regression TRAIN
Decision Tree TRAIN

Accuracy


Accuracy
0.8
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
0.8
1.0
Cutoff
Data Role
Logistic Regression TEST
Decision Tree TEST
Logistic Regression TRAIN
Decision Tree TRAIN

F1 Score


F1 Score
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Cutoff
Data Role
Logistic Regression TEST
Decision Tree TEST
Logistic Regression TRAIN
Decision Tree TRAIN

Fit Statistics

Statistics Label	TRAIN_Logistic_Regression	TEST_Logistic_Regression	TRAIN_Decision_Tree
Area Under ROC	0.6994	0.7028	0.8116
Average Squared Error	0.2054	0.2056	0.1736
Divisor for ASE	2,467	1,056	2,467
Formatted Partition	1	2	1
Gamma	0.5643	0.5661	0.9960
Gini Coefficient	0.3989	0.4055	0.6232
KS (Youden)	0.3323	0.3540	0.4284
KS at Default Cutoff	0.2460	0.2018	0.3416
KS Cutoff	0.3600	0.3600	0.0100
Misclassification Rate	0.5334	0.5275	0.4682
Misclassification Rate (Event)	0.2983	0.3201	0.2359
Misclassification Rate at KS Cutoff (Event)	0.3482	0.3456	0.3672
Multi-Class Log Loss	1.0237	1.0232	0.8342
Number of Observations	2,467	1,056	2,467
Partition Indicator	1	2	1
Root Average Squared Error	0.4532	0.4534	0.4167
Target Name	acci_severity	acci_severity	acci_severity
Tau	0.1833	0.1866	0.2864

TEST_Decision_Tree
0.8205
0.1768

TEST_Decision_Tree
1,056
2
0.9913
0.6411
0.4381
0.3633
0.0100
0.4782
0.2292
0.3608
0.8457
1,056
2
0.4205
acci_severity
0.2950

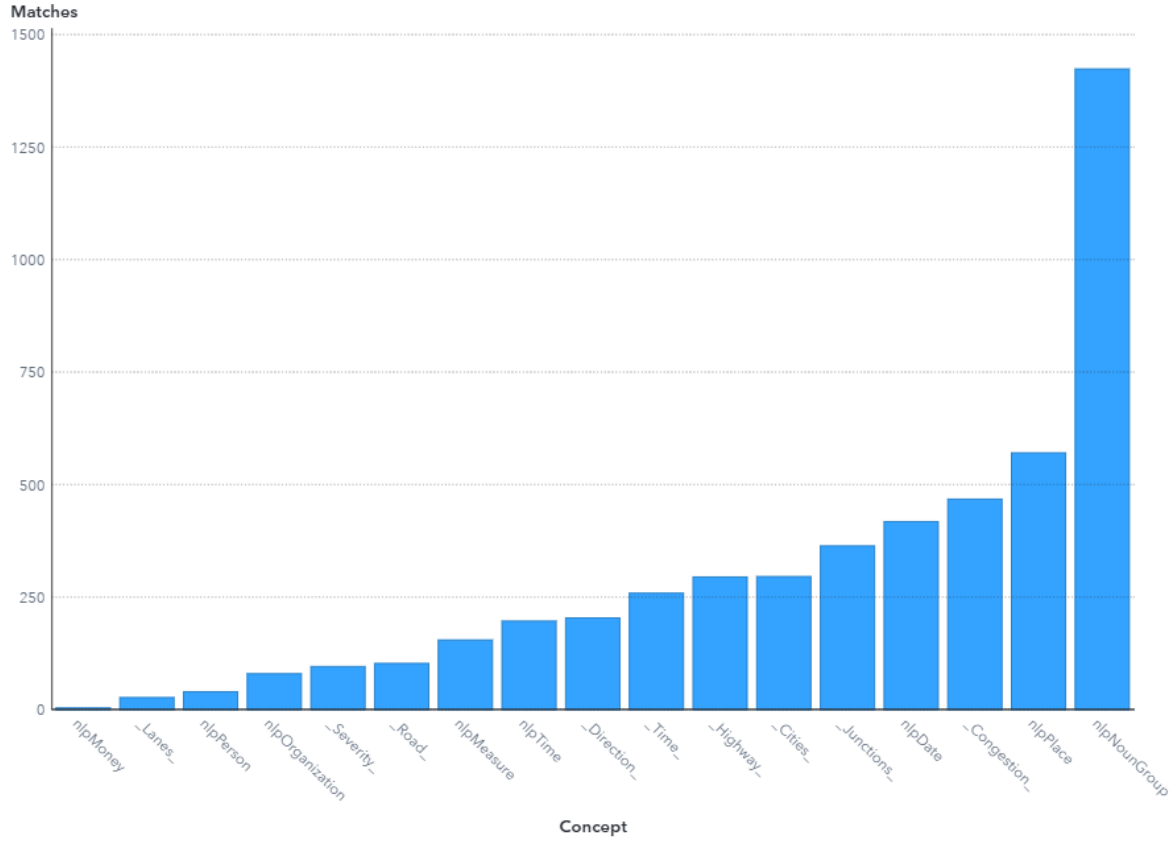
tweets3

"Concepts" Results

by: ta01468@surrey.ac.uk

Contents

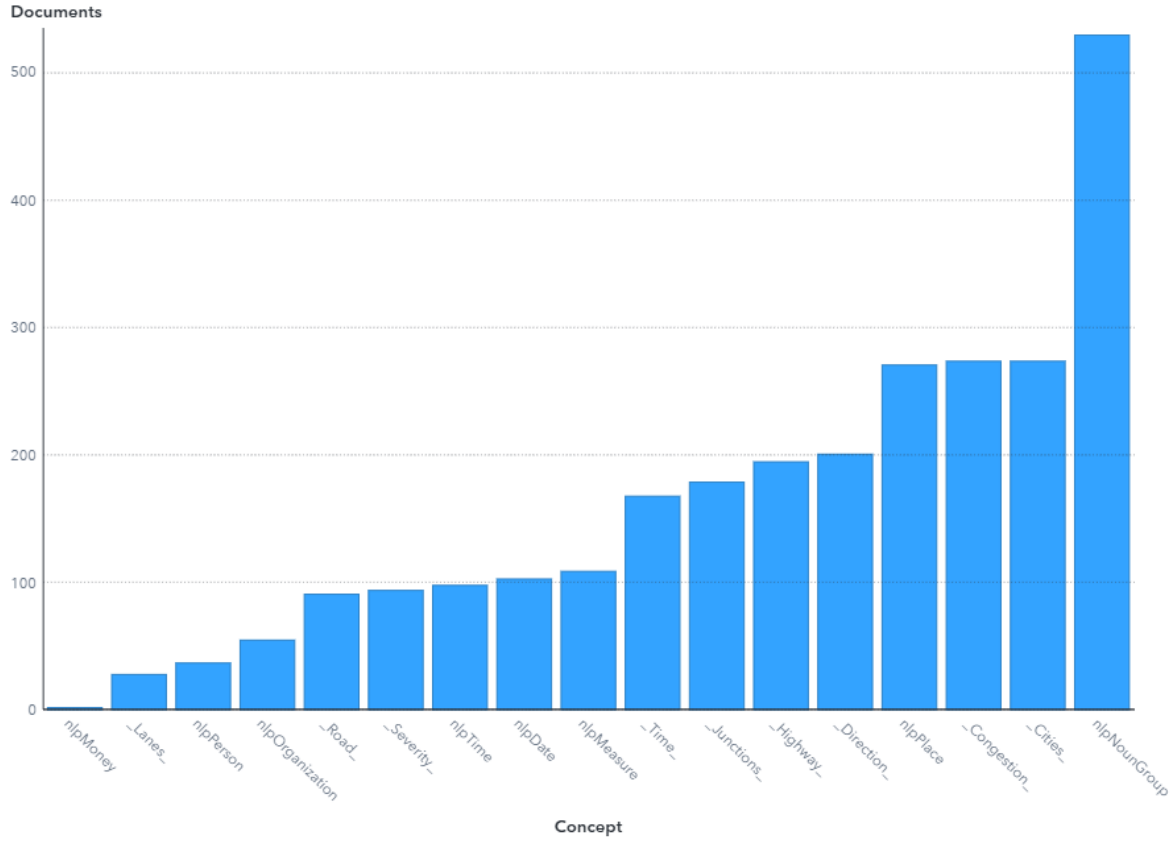
Number of Matches Per Concept



The Number of Matches per Concept report depicts how useful each concept is for finding information in the data generally. The top matching concept in this data, nlpNounGroup, has 1,425 matches, while the least matched concept, nlpMoney, has 5 matches.

This information indicates how closely each concept, and the data in the documents are aligned. Many matches show that a concept is well-defined to extract information from the data set. Fewer matches for a concept indicate that either the data is not appropriate for the concept or that the concept needs further definition by adding or refining rules.

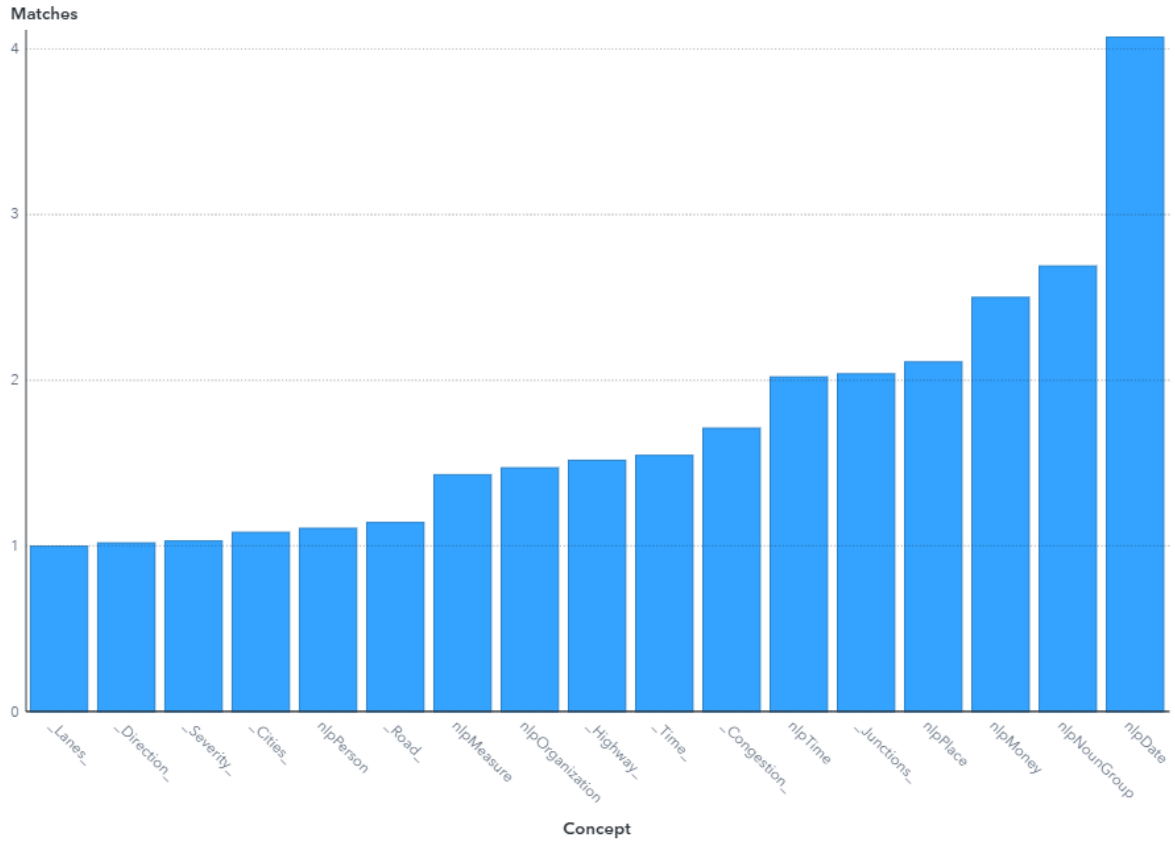
Number of Documents Per Concept



The Number of Documents per Concept report depicts how well each concept covers the documents in the data set. In this data set, the two concepts with the greatest coverage are nlpNounGroup with matches spanning across 530 documents (88.63% of the total documents), and _Congestion_ with matches spanning across 274 documents (45.82% of total). The concepts with very light coverage (less than 5%) across documents in this data set are: _Lanes_, nlpMoney.

This information indicates how broad each concept is in terms of how many documents it matches. In a project that is expected to cover all the documents with each concept or a subset of concepts, this report can be used to gauge how close the model is to that goal and which concepts are more complete in their coverage.

Average Number of Matches Per Document



The Average Number of Matches per Document report depicts the amount of information extraction performed by the concept inside each document where it matches. In this data set, for example, when nlpNounGroup matches in a document, it matches 2.69 times on average.

The calculation of the average does not include documents where the match count is zero for that concept. This report shows when a concept is getting the desired depth to help with prioritization during development of concepts within the model.

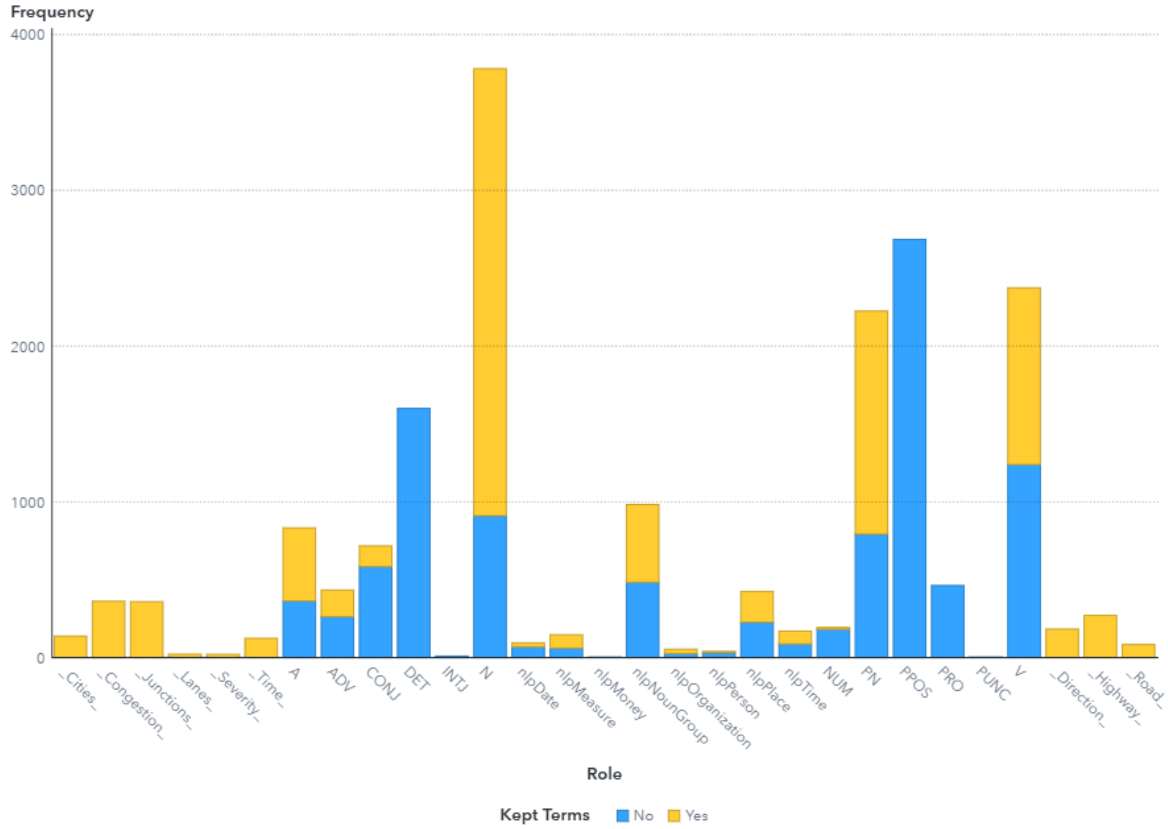
tweets3

"Text Parsing" Results

by: ta01468@surrey.ac.uk

Contents

Role by Frequency



The Role by Frequency report is a visual summary of the terms data. Parsing assigns role labels to each term in the entire data set. The height of the bars in this report indicates the frequency for each type of role. In this data set, the most frequently occurring roles are N (noun) and PPOS (preposition/postposition). Together they represent 34.25% of the terms. Note that both bare numbers and most types of punctuation (except some symbols) are removed by default from the terms list and are not reported here. To work with the table of values further, download the data.

For each bar, the areas defined by the two colors represent the proportion of kept or dropped terms with the given role. A downstream Topics node will use only the kept terms to determine topics across the data set. In this data set, 53.82% of the terms that are found in the data have been dropped from consideration in building topics. The dropped terms were removed because of their presence in a stop list or because they do not meet the frequency threshold set by the ‘minimum number of documents’ parameter. Additional terms can be dropped in the interactive view or by adding them to the applied stop list.

Parsing identifies terms through a sequence of NLP (natural language processing) steps, including tokenization, multiword identification, lemmatization, part-of-speech tagging, and noun group extraction. Synonym lists and misspelling detection can be

added to the analysis to further group term variants under a single parent term. Concept extraction may also be applied using a preceding Concepts node.

Each resulting role is either a part-of-speech (content or function word) or a concept. Parts-of-speech include labels such as N (noun), CONJ (conjunction), PN (proper noun), ADV (adverb), NUM (numeric), PUNC (punctuation), and so on. Concepts may include noun groups (nlpNounGroup) and those defined and passed forward from a preceding Concepts node in the pipeline.

Descriptive Statistics

Measure	Terms in a Sentence	Terms in a Document
Minimum	1	1
Maximum	52	66
Mean	14.5624	31.6087

The Descriptive Statistics table records patterns that are found across the data set. The average length of sentences by count of terms in this data set is 14.56. The range of sentence length is 1 - 52 terms. The average length of documents by a count of terms in this data set is 31.61. The range of document length is 1 - 66 terms.

The information in this chart can identify unexpected data characteristics that may need to be investigated. The information presented in the table is a summary view only. For more detailed information or to compare data sets, run the profileText action.

tweets3

"Sentiment" Results

by: ta01468@surrey.ac.uk

Contents

Sentiment Score Code

```
/*****************************************************************
* SAS Visual Text Analytics
* Sentiment Score Code
*
* Modify the following macro variables to match your needs.
*****************************************************************/

/* specifies CAS library information for the CAS table that you would like to score.
You must modify the value to provide the name of the library that contains the table
to be scored. */
%let input_caslib_name = "{input_caslib_name}";

/* specifies the CAS table you would like to score. You must modify the value to
provide the name of the input table, such as "MyTable". Do not include an extension.
*/
%let input_table_name = "{input_cas_table_name}";

/* specifies the column in the CAS table that contains a unique document identifier.
You must modify the value to provide the name of the document identifer column in
the table. */
%let key_column = "{doc_id_column_name}";

/* specifies the column in the CAS table that contains the text data to score. You
must modify the value to provide the name of the text column in the table. */
%let document_column = "{text_column_name}";

/* specifies the CAS library to write the score output tables. You must modify the
value to provide the name of the library that will contain the output tables that the
score code produces. */
%let output_caslib_name = "{output_caslib_name}";

/* specifies the sentiment output CAS table to produce */
%let output_sentiment_table_name = "out_sentiment";

/* specifies the matches output CAS table to produce */
%let output_matches_table_name = "out_sent_matches";

/* specifies the features output CAS table to produce */
%let output_features_table_name = "out_sent_features";
```

```
/* specifies the language of the associated SAS Visual Text Analytics project. This
should be set automatically to the language you selected when you created your
project. */
%let language = "ENGLISH";

/* specifies the hostname for the CAS server. This should be set automatically to the
host for the associated SAS Visual Text Analytics project. */
%let cas_server_hostname = "sas-cas-server-default-client";

/* specifies the port for the CAS server. This should be set automatically to the host
for the associated SAS Visual Text Analytics project. */
%let cas_server_port = 5570;

/* creates a session */
cas sascas1 host=&cas_server_hostname port=&cas_server_port;
libname sascas1 cas sessref=sascas1 datalimit=all;

/* calls the scoring action */
proc cas;
session sascas1;
loadactionset "sentimentAnalysis";

action applySent;
param
table={caslib=&input_caslib_name, name=&input_table_name}
docId=&key_column
text=&document_column
language=&language
casOut={caslib=&output_caslib_name, name=&output_sentiment_table_name,
replace=TRUE}
matchOut={caslib=&output_caslib_name, name=&output_matches_table_name,
replace=TRUE}
featureOut={caslib=&output_caslib_name, name=&output_features_table_name,
replace=TRUE}
;
run;
quit;
```

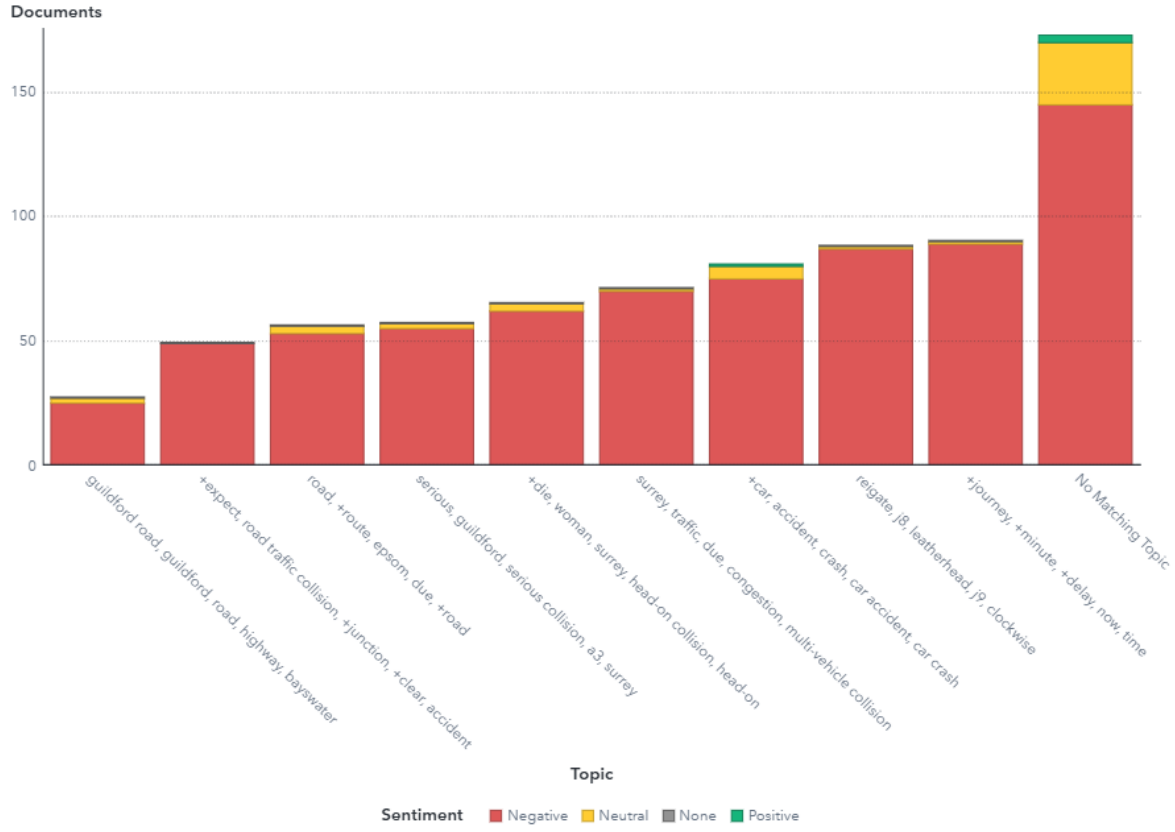
tweets3

"Topics" Results

by: ta01468@surrey.ac.uk

Contents

Number of Documents Per Topic



The Number of Documents per Topic report is a visual summary of the topics in the data. 9 of the bars represent the number of topics identified. Bar 10 in the chart represents the documents that have not been assigned to any topic. The height of each bar represents the number of documents placed in each topic. Documents may be placed into more than a single topic.

Applying this set of topics to this data set results in 173 (28.93%) of the documents not being placed in any topic. If more coverage of the data is desired, the cutoff for documents and/or terms can be lowered, or more topics can be added, or existing topics expanded in scope.

If a Sentiment node is placed before the Topics node in the pipeline, each bar representing a topic is split between positive, neutral, negative, and no sentiment by color.

tweets3

"Categories" Results

by: ta01468@surrey.ac.uk

Contents

Categories Score Code

```
/*****************************************************************
* SAS Visual Text Analytics
* Categories Score Code
*
* Modify the following macro variables to match your needs.
*****************************************************************/

/* specifies CAS library information for the CAS table that you would like to score.
You must modify the value to provide the name of the library that contains the table
to be scored. */
%let input_caslib_name = "{input_caslib_name}";

/* specifies the CAS table you would like to score. You must modify the value to
provide the name of the input table, such as "MyTable". Do not include an extension.
*/
%let input_table_name = "{input_cas_table_name}";

/* specifies the column in the CAS table that contains a unique document identifier.
You must modify the value to provide the name of the document identifer column in
the table. */
%let key_column = "{doc_id_column_name}";

/* specifies the column in the CAS table that contains the text data to score. You
must modify the value to provide the name of the text column in the table. */
%let document_column = "{text_column_name}";

/* specifies the CAS library to write the score output tables. You must modify the
value to provide the name of the library that will contain the output tables that the
score code produces. */
%let output_caslib_name = "{output_caslib_name}";

/* specfies the categories output CAS table to produce */
%let output_categories_table_name = "out_categories";

/* specifies the matches output CAS table to produce */
%let output_matches_table_name = "out_matches";

/* specifies the modeling ready output CAS table to produce */
%let output_modeling_ready_table_name = "out_modeling_ready";
```

```
/* specifies the CAS library information for the mco binary table. This should be set
automatically to the CAS library for the associated SAS Visual Text Analytics project.
*/
%let mco_binary_caslib = "Analytics_Project_abdf61d7-63d1-4cd9-
a0bd-7a51e1deb14a";

/* specifies the CAS table name of the mco binary table. This should be set
automatically to the Categories node model table for the associated SAS Visual Text
Analytics project. */
%let mco_binary_table_name = "f6b7b8df-ac3d-4114-935c-
ebc5a5a06c6c_CATEGORY_BINARY";

/* specifies the hostname for the CAS server. This should be set automatically to the
host for the associated SAS Visual Text Analytics project. */
%let cas_server_hostname = "sas-cas-server-default-client";

/* specifies the port for the CAS server. This should be set automatically to the host
for the associated SAS Visual Text Analytics project. */
%let cas_server_port = 5570;

/* creates a session */
cas sascas1 host=&cas_server_hostname port=&cas_server_port
uuidmac=sascas1_uuid;
libname sascas1 cas sessref=sascas1 datalimit=all;

/* calls the scoring action */
proc cas;
session sascas1;
loadactionset "textRuleScore";

action applyCategory;
param
model={caslib=&mco_binary_caslib, name=&mco_binary_table_name}
table={caslib=&input_caslib_name, name=&input_table_name}
docId=&key_column
text=&document_column
casOut={caslib=&output_caslib_name, name=&output_categories_table_name,
replace=TRUE}
matchOut={caslib=&output_caslib_name, name=&output_matches_table_name,
replace=TRUE}
modelOut={caslib=&output_caslib_name,
name=&output_modeling_ready_table_name, replace=TRUE}
;
```

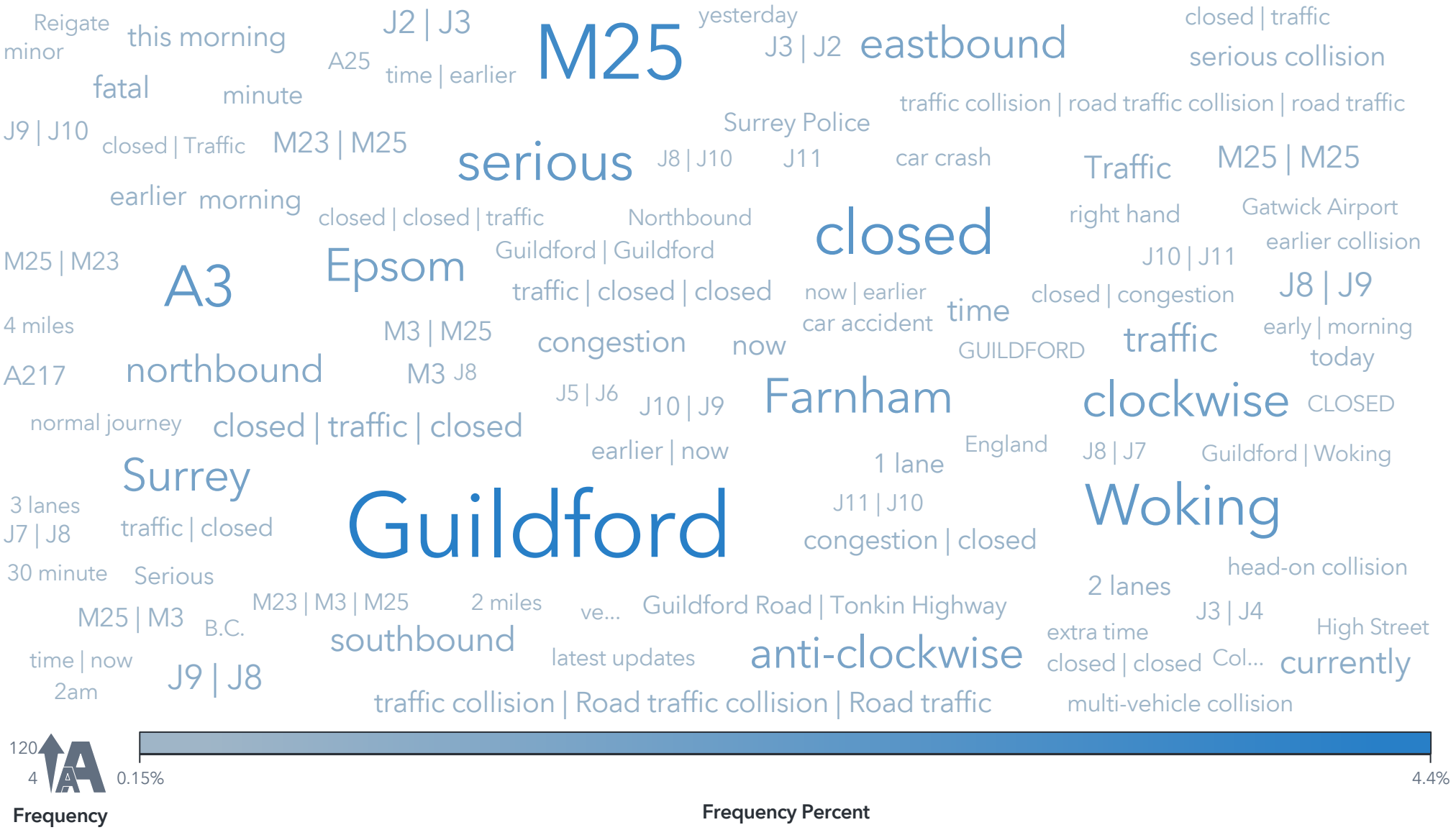
```
run;
quit;
```

Tweet_Graphs_Final

Creation Date: Thursday, January 9, 2025, 10:04:47 PM

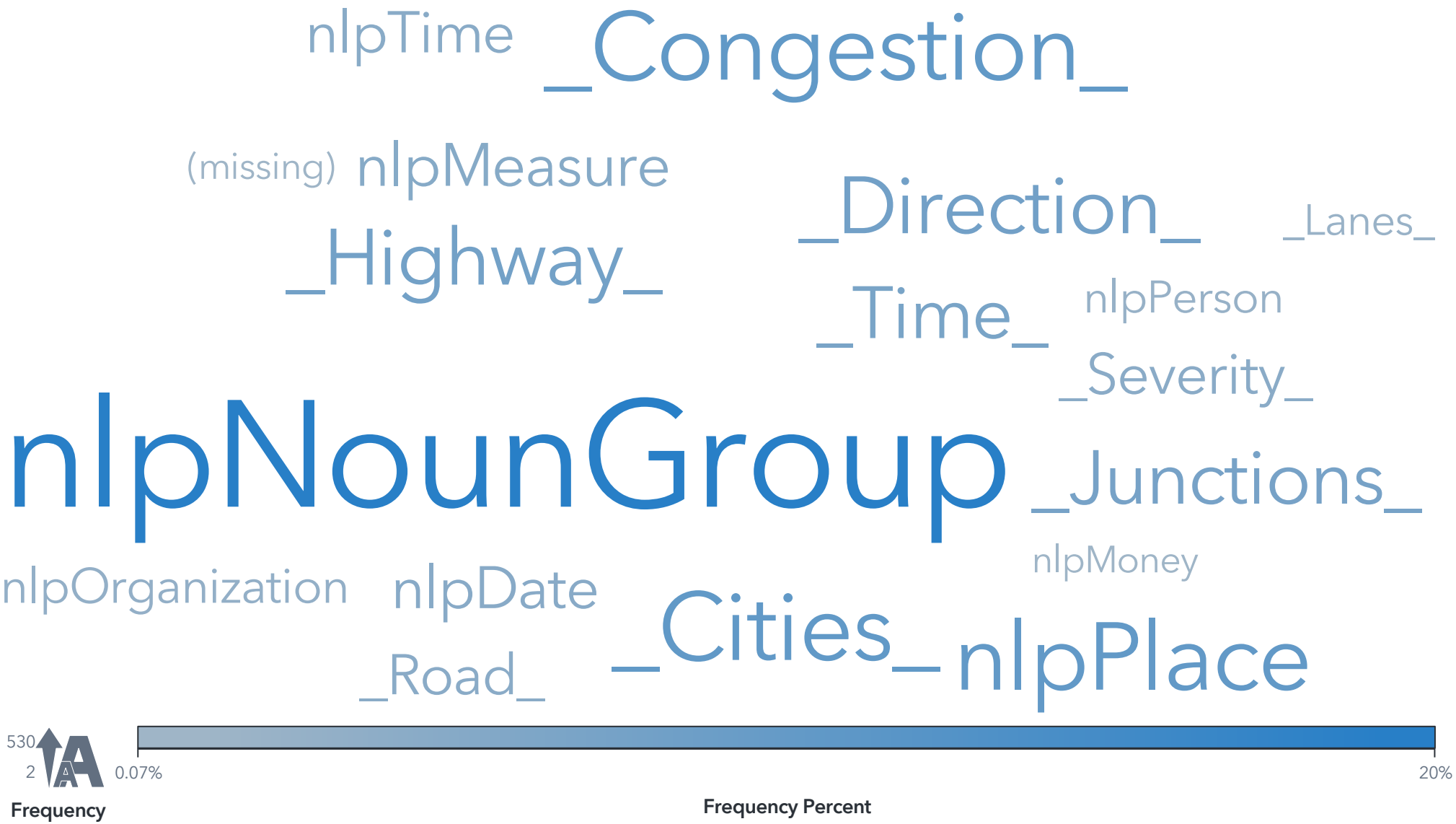
Author: ta01468@surrey.ac.uk

Frequency, Frequency Percent by Keywords



A1.1

Frequency, Frequency Percent by Concept Name



Frequency, Frequency Percent by Text

A2.1

Appendix

A1.1 Frequency, Frequency Percent by Keywords

Warnings: Only 100 rows of the data appear.

A2.1 Frequency, Frequency Percent by Text

Warnings: Only 100 rows of the data appear.