Meenu Ravi

MSCA31010: Linear & Non-Linear Models

Winter 2021 Assignment 2

You are asked to train a binary logistic regression model on the claim_history.csv. Your model will predict the likelihood of filing more than one claim in one unit of exposure. You will first calculate the Frequency variable by dividing the CLM_COUNT by EXPOSURE. Next, you will create a binary target variable that determines if the Frequency is strictly greater than one (i.e., the Event).

You will use MSTATUS, CAR_TYPE, REVOKED, and URBANICITY as the categorical predictors, and CAR_AGE, MVR_PTS, TIF, and TRAVTIME as the interval predictors. Your goal is to train a model that has just the right set of predictors.

You must perform the calculations without calling any special libraries (e.g., scikit-learn or statsmodels). The standard libraries such as numpy and pandas are allowed. You need to drop all missing values (i.e., NaN) of all the predictors and the target variable before training your model.

Question 1 (25 points)

Before you train the model, you want to explore the predictors.

a) (15 points) For each predictor, generate a line chart that shows the odds of the Event by the predictor's unique values. The predictor's unique values are displayed in ascending lexical order.

The resulting line charts for the odds of the event by the predictor's unique values are displayed at the end of this document

b) (10 points) Also, calculate the ratio of the maximum odds value to the minimum odds value. If the minimum odds value is zero, then the ratio is infinity. Based on the ratio, please provide us your opinions of whether the final model will include that predictor.

Variables	ratio of max odds value to the min odds	Will final model include Variable?	Explanation

Mstatus	1.758847126	possible	the ratio in second column is greater than 1 but only slightly greater
Car_type	2.443824228	very likely	the ratio in second column is greater than 1- a lot greater
Revoked	2.561723628	very likely	the ratio in second column is greater than 1- a lot greater
Urbanicity	6.625049179	very likely	the ratio in second column is greater than 1- a lot greater
Car_age	Inf	Could be included	We are not able to make a clear prediction because the ratios are infinity. These are interval variables so their values are unpredictable
Mvr_pts	Inf	Could be included	We are not able to make a clear prediction because the ratios are infinity. These are interval variables so their values are unpredictable
Tif	Inf	Could be included	We are not able to make a clear prediction because the ratios are infinity. These are interval variables so their values are unpredictable
Travtime	Inf	Could be included	We are not able to make a clear prediction because the ratios are infinity. These are interval variables so their values are unpredictable

Question 2 (40 points)

Enter the predictors into your model using Forward Selection. The Entry Threshold is 0.05.

a) (20 points). Please provide a detailed report of the Forward Selection. However, you do not need to show steps such as 1.1. The report should include (1) the predictor entered, (2) the number of free parameters, (3) the log-likelihood value, (4) the Deviance Chi-squares statistic, (5) the Deviance Degree of Freedom, and (6) the Chi-square significance.

Step	Parameter Entered	Free Parameters	Log Likelihood Value	Deviance Chi-squares	Deviance DoF	Chi-square sig
0	intercept	1	-5413.971792	-	-	-
1	URBANICITY	2	-5124.8897	578.164184	1	9.41E-128

2	MVR_PTS	3	-4969.266156	311.2470883	1	1.17E-69
3	CAR_AGE	4	-4884.769645	168.9930208	1	1.23E-38
4	MSTATUS	5	-4811.458945	146.6214014	1	9.50E-34
5	REVOKED	6	-4744.787224	133.3434411	1	7.60E-31
6	CAR_TYPE	11	-4673.497664	142.5791196	5	5.06E-29
7	TRAVTIME	12	-4632.903799	81.18773184	1	2.05E-19
8	TIF	13	-4604.172367	57.46286215	1	3.44E-14

*full attached at end

b) (5 points). Which predictors does your final model contain?

My final model contains all the predictors MSTATUS, CAR_TYPE, REVOKED, and URBANICITY, CAR_AGE, MVR_PTS, TIF, and TRAVTIME.

c) (5 points). What are the aliased parameters in your final model? Please list the predictor's name and the aliased categories.

Aliased parameters	Category	Full Name
CAR_TYPE	Van	CAR_TYPE_Van
REVOKED	YES	REVOKED_YES
MSTATUS	YES	MSTATUS_YES
URBANICITY	Highly Urban/Urban	URBANICITY_Highly Urban/Urban

d) (5 points). How many non-aliased parameters are in your final model?

There are 12 non aliased parameters in the final model.

e) (5 points). Please show a table of the complete set of parameters of your final model (including the aliased parameters). Besides the parameter estimates, please also include the exponentiated estimates (i.e., apply the exp() function on the parameter estimates).

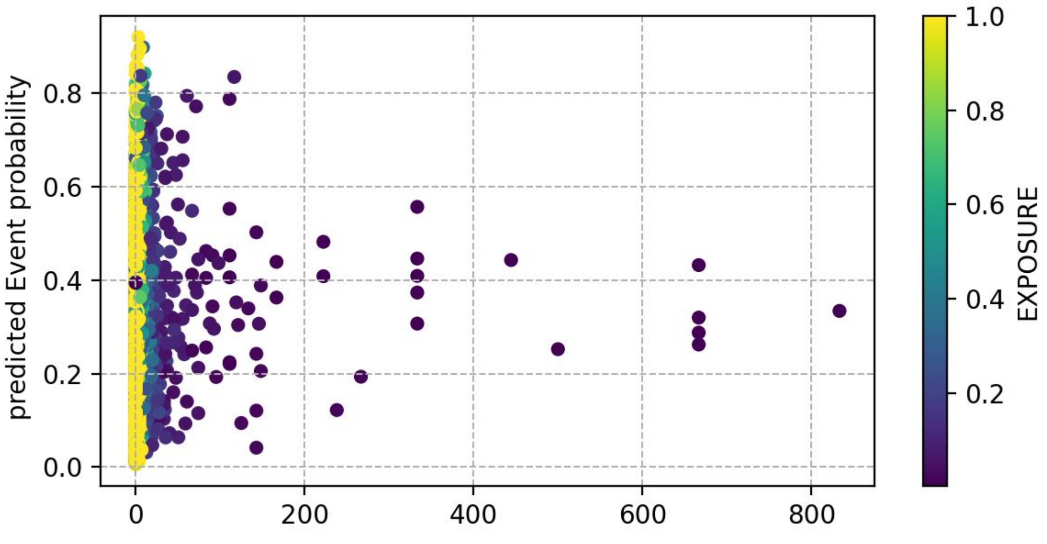
Parameters	Parameters Estimates	Estimates Exponentiated
Intercept	-0.5991579911	0.549273935
CAR_TYPE_Minivan	-0.479322972	0.6192024673
CAR_TYPE_Panel Truck	0.07441589842	1.077254741
CAR_TYPE_Pickup	0.2536612298	1.288735145
CAR_TYPE_SUV	0.1643072518	1.178576379

CAR_TYPE_Sports Car	0.4688522784	1.598158899
CAR_TYPE_Van	0	1
REVOKED_No	-0.7838057224	0.4566647607
REVOKED_Yes	0	1
MSTATUS_No	0.6369460891	1.89069803
MSTATUS_Yes	0	1
URBANICITY_Highly Rural/	-2.079242571	0.1250248738
URBANICITY_Highly Urban/Urban	0	1
MVR_PTS	0.1637937382	1.17797132
CAR_AGE	-0.05888064638	0.9428192913
TRAVTIME	0.01494425932	1.015056483
TIF	-0.04885915695	0.9523152472

Question 3 (20 points)

You will visually assess your final model in Question 2. Please color-code the markers according to the Exposure value. Also, please briefly comment on the graphs.

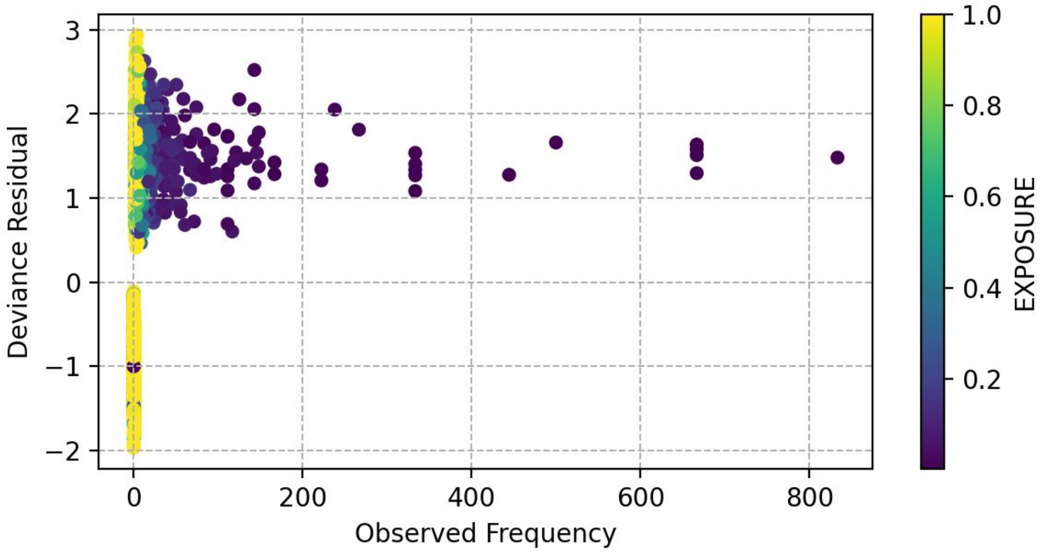
a) (10 points). Please plot the predicted Event probability versus the observed Frequency.



Observed Frequency

As the observed frequency increases, the exposure decreases and the predicted probability seems to increase slightly. As exposure increases, the observed frequency is very low, almost close to 0 and the number of claims increases.

b) (10 points). Please plot the Deviance residuals versus the observed Frequency.



This graph shows that the deviance residual is less than 0 for lowers observed frequencies. For these customers, the exposure is higher, so they are with us for a longer time. As the deviance residual increases, we see more observed frequencies because more people are filing claims, and their exposure is lower.

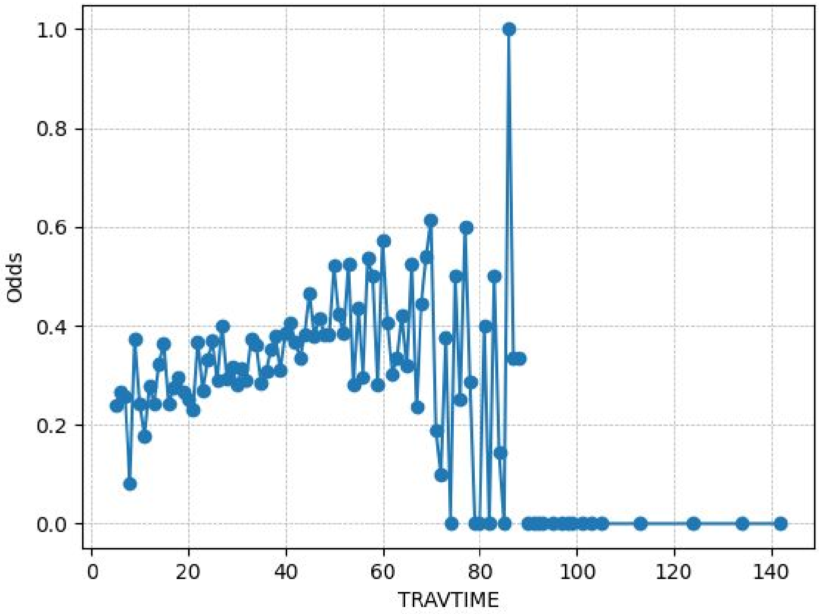
Question 4 (15 points)

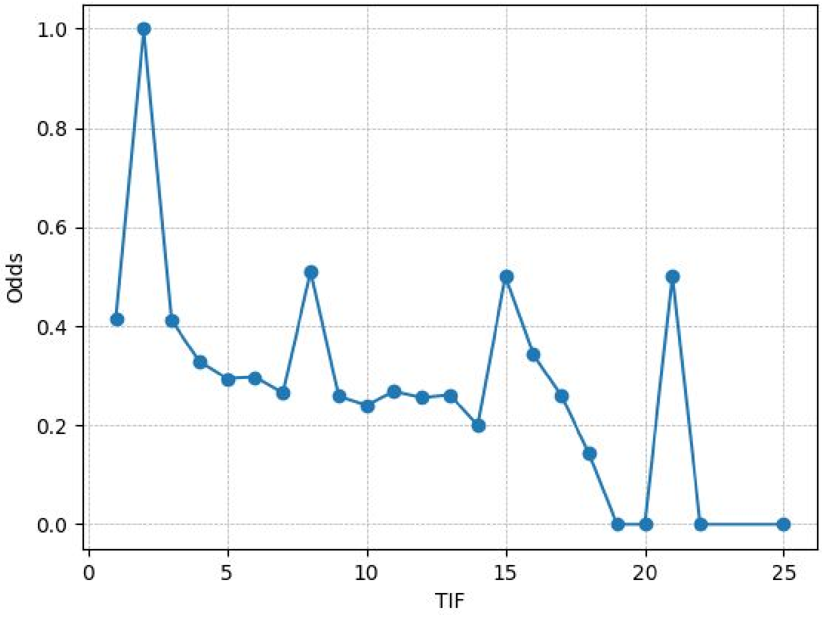
You will calculate the Accuracy metric to assess your final model in Question 2. If the predicted Event probability of an observation is greater than or equal to 0.25, then you will classify that observation as the Event (i.e., filing more than one claim per unit exposure). An observation is correctly classified if the predicted target value equals the observed target value. The Accuracy metric is the proportion of observations that are correctly classified.

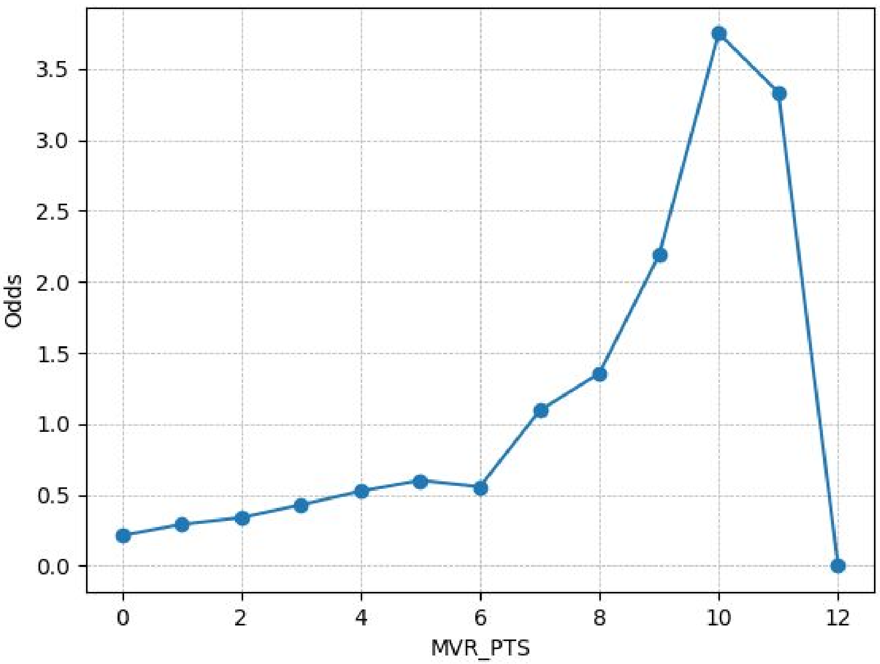
The accuracy metric is 0.675. This shows that our model is not very good. It is accuracy a little over half the time.

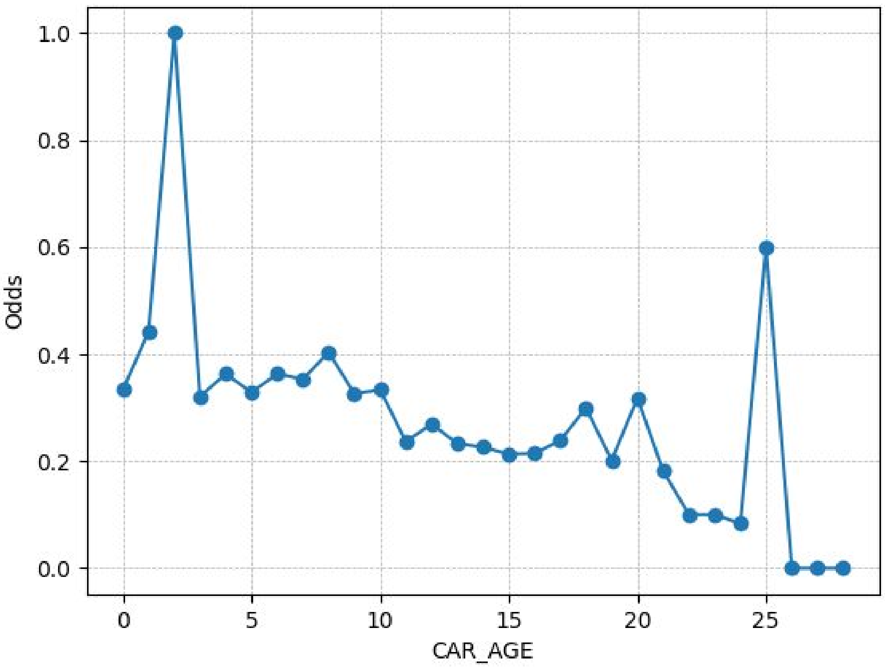
Accuracy metric	0.6753260195

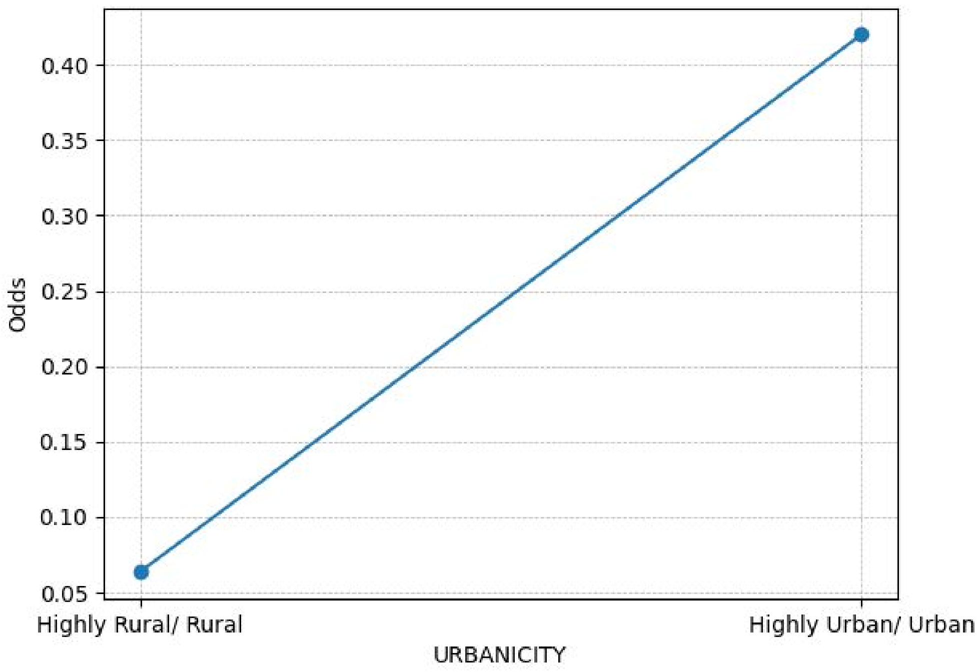
Line charts for the odds of the event by the predictor's unique values:

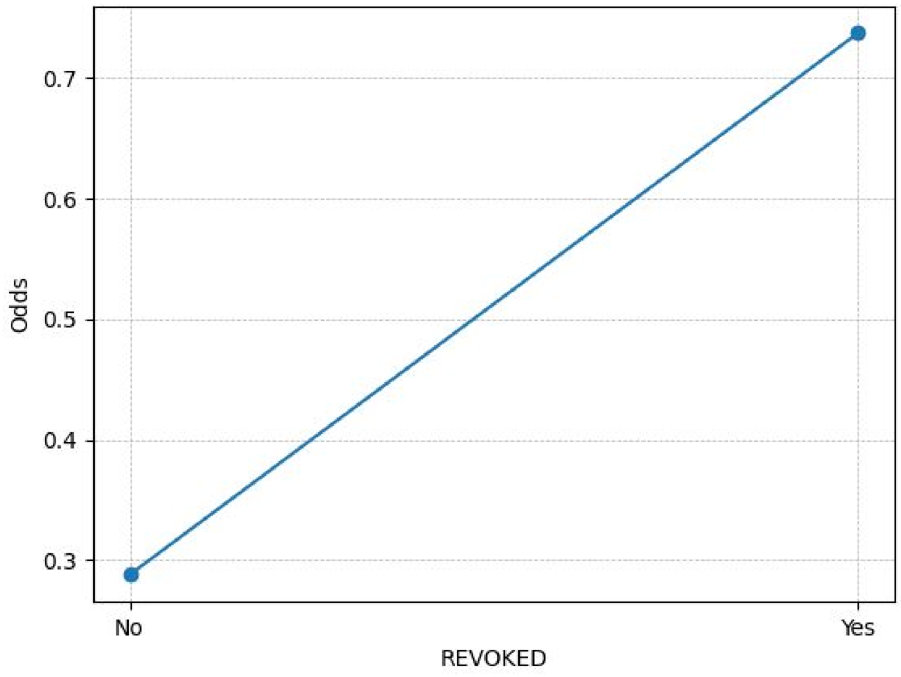


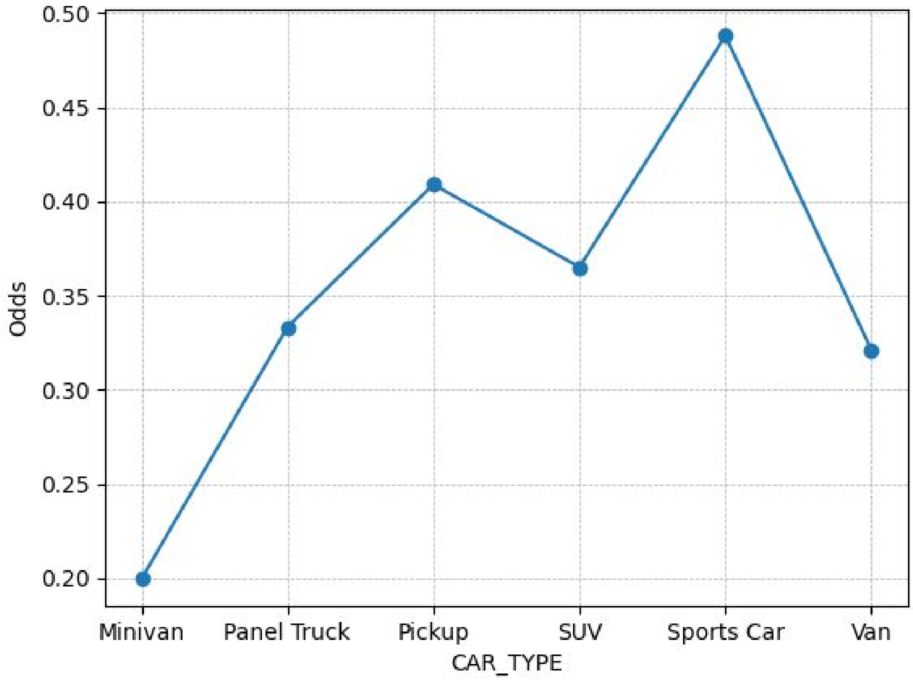


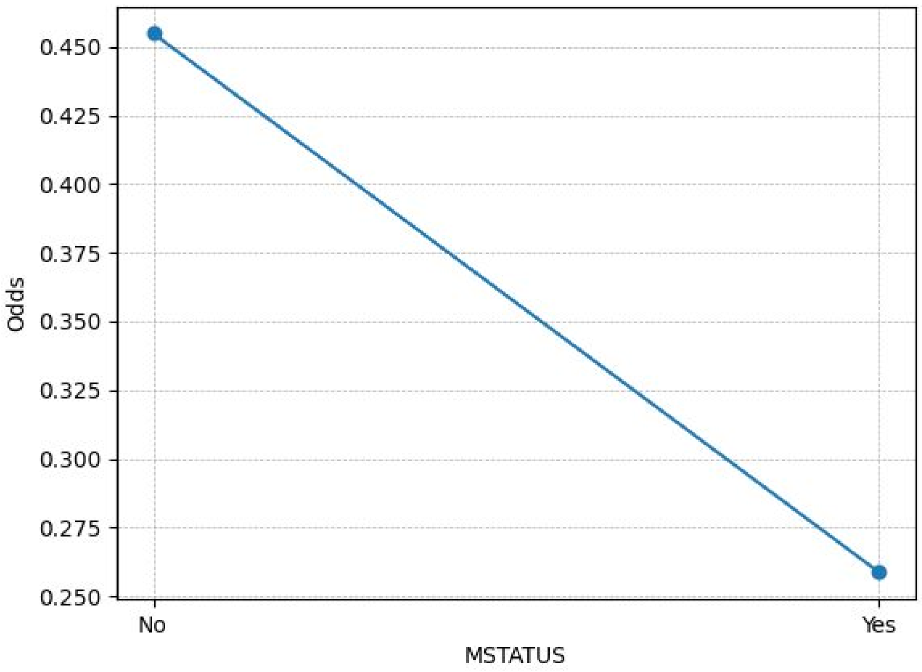

Odds
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
10
12
MVR_PTS


Odds
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
CAR_AGE


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
Odds
Highly Rural/ Rural
Highly Urban/ Urban
URBANICITY


0.7
0.6
0.5
0.4
0.3
Odds
No
Yes
REVOKED


Odds
0.50
0.45
0.40
0.35
0.30
0.25
0.20
Minivan
Panel Truck
Pickup
SUV
Sports Car
Van
CAR_TYPE


Odds
0.450
0.425
0.400
0.375
0.350
0.325
0.300
0.275
0.250
No
Yes
MSTATUS

Full forward charts:

Step 0	Model Parameters	Free Parameters	Log Liklihood Value	Deviance Chi-squares	Deviance DoF	Chi-square sig
0	intercept	1	-5413.971792	-	-	-
1	MSTATUS	2	-5343.629436	140.6847115	1	1.89E-32
2	CAR_TYPE	6	-5336.549961	154.8436629	5	1.24E-31
3	REVOKED	2	-5312.143428	203.6567276	1	3.33E-46
4	URBANICITY	2	-5124.8897	578.164184	1	9.41E-128
5	CAR_AGE	2	-5362.824193	102.295199	1	4.78E-24
6	MVR_PTS	2	-5199.195271	429.5530422	1	2.03E-95
7	TIF	2	-5382.987153	61.96927827	1	3.49E-15
8	TRAVTIME	2	-5400.307781	27.3280225	1	1.72E-07
Enter Urbanicity						
Step 1	Model Parameters	Free Parameters	Log Liklihood Value	Deviance Chi-squares	Deviance DoF	Chi-square sig
0	intercept + URBANICITY	2	-5124.8897	-	-	-
1	MSTATUS	3	-5051.389707	146.9999867	1	7.85E-34
2	CAR_TYPE	7	-5031.601961	186.5754777	5	2.11E-38

3	REVOKED	3	-5045.639347	158.500706	1	2.41E-36
4	CAR_AGE	3	-5026.067614	197.6441723	1	6.82E-45
5	MVR_PTS	3	-4969.266156	311.2470883	1	1.17E-69
6	TIF	3	-5091.330652	67.11809646	1	2.56E-16
7	TRAVTIME	3	-5081.929617	85.92016564	1	1.87E-20
Enter MVR_PTS						
Step 2	Model Parameters	Free Parameters	Log Liklihood Value	Deviance Chi-squares	Deviance DoF	Chi-square sig
0	intercept+URBANICITY+MVR_PTS	3	-4969.266156	-	-	-
1	MSTATUS	4	-4905.16175	128.2088125	1	1.01E-29
2	CAR_TYPE	8	-4890.36759	157.7971308	5	2.92E-32
3	REVOKED	4	-4894.769004	148.9943037	1	2.88E-34
4	CAR_AGE	4	-4884.769645	168.9930208	1	1.23E-38
5	TIF	4	-4939.303592	59.92512808	1	9.85E-15
6	TRAVTIME	4	-4930.999899	76.53251299	1	2.17E-18
Enter CAR_AGE						
Step 3	Model Parameters	Free Parameters	Log Liklihood Value	Deviance Chi-squares	Deviance DoF	Chi-square sig

0	intercept+URBANICITY+MVR_PTS+CAR_AGE	4	-4884.769645	-	-	-
1	MSTATUS	5	-4811.458945	146.6214014	1	9.50E-34
2	CAR_TYPE	9	-4814.393387	140.7525177	5	1.24E-28
3	REVOKED	5	-4813.622569	142.2941521	1	8.39E-33
4	TIF	5	-4855.390285	58.75872062	1	1.78E-14
5	TRAVTIME	5	-4846.280281	76.97872927	1	1.73E-18
Enter MSTATUS						
Step 4	Model Parameters	Free Parameters	Log Liklihood Value	Deviance Chi-squares	Deviance DoF	Chi-square sig
0	intercept+URBANICITY+MVR_PTS+CAR_AGE+MSTATUS	5	-4811.458945	-	-	-
1	CAR_TYPE	10	-4738.290057	146.3377759	5	8.03E-30
2	REVOKED	6	-4744.787224	133.3434411	1	7.60E-31
3	TIF	6	-4780.363702	62.19048591	1	3.12E-15
4	TRAVTIME	6	-4770.963084	80.99172206	1	2.27E-19
Enter REVOKED						

Step 5	Model Parameters	Free Parameters	Log Liklihood Value	Deviance Chi-squares	Deviance DoF	Chi-square sig
0	intercept+URBANICITY+ MVR_PTS+ CAR_AGE+ MSTATUS+ REVOKED	6	-4744.787224	-	-	-
1	CAR_TYPE	11	-4673.497664	142.5791196	5	5.06E-29
2	TIF	7	-4716.50091	56.57262857	1	5.42E-14
3	TRAVTIME	7	-4704.833956	79.9065363	1	3.93E-19
Enter CAR_TYPE						
Step 6	Model Parameters	Free Parameters	Log Liklihood Value	Deviance Chi-squares	Deviance DoF	Chi-square sig
0	intercept+URBANICITY+ MVR_PTS+ CAR_AGE+ MSTATUS+ REVOKED+ CAR_TYPE	11	-4673.497664	-	-	-
1	TIF	12	-4644.128022	58.73928464	1	1.80E-14
2	TRAVTIME	12	-4632.903799	81.18773184	1	2.05E-19
Enter TRAVTIME						

Step 7	Model Parameters	Free Parameters	Log Liklihood Value	Deviance Chi-squares	Deviance DoF	Chi-square sig
0	intercept+URBANICITY+MVR_PTS+CAR_AGE+MSTATUS+REVOKED+CAR_TYPE+TRAVTIME	12	-4632.903799	-	-	-
1	TIF	13	-4604.172367	57.46286215	1	3.44E-14