





































JMP Solution

Steps followed

1. Built an initial model without any transformation, including all variables
2. Checked for ROC curve, Confusion matrix, sensitivity and specificity
3. Compared the results by varying forward, backward stepwise; AIC/BIC criteria (All these gave very similar results)
4. Plotted marginal plots to check for further significance and validity of variables
5. Chose a few variables could be improved based on the skewness of above plot. The variables chosen are our, over, order, mail, report, your, C!

Source	LogWorth		PValue
George	54.459		0.00000
Hp	27.734		0.00000
Remove	23.418		0.00000
Free	21.835		0.00000
C\$	19.896		0.00000
Edu	15.687		0.00000
Meeting	12.148		0.00000
re:	11.339		0.00000
W_000	9.928		0.00000
Our	9.677		0.00000
C!	7.845		0.00000
Cs	6.557		0.00000
Business	5.667		0.00000
CAP_tot	5.464		0.00000
Your	5.418		0.00000
Internet	5.142		0.00001
Project	4.811		0.00002
C;	4.512		0.00003
Conferenc	4.342		0.00005
e			
CAP_long	4.341		0.00005
Over	4.151		0.00007
Money	3.744		0.00018
Credit	3.463		0.00034
Data	3.040		0.00091
Technolog	2.785		0.00164
y			
Lab	2.661		0.00218
Hpl	2.355		0.00441
W_85	2.332		0.00465
W_650	2.195		0.00639
Order	2.158		0.00694
Pm	2.111		0.00774
C#	2.083		0.00826
You	2.077		0.00837
Original	1.458		0.03486
Table	1.451		0.03538
Will	1.314		0.04858

[illegible]

Effect Likelihood Ratio Tests

Source	Nparm	DF	ChiSquare	Prob>ChiSq
our	1	1	40.3670939	<.0001*
over	1	1	15.7942588	<.0001*
remove	1	1	102.739114	<.0001*
internet	1	1	20.1363553	<.0001*
order	1	1	7.28715546	0.0069*
mail	1	1	3.42460983	0.0642
receive	1	1	3.114538	0.0776
will	1	1	3.88996583	0.0486*
people	1	1	0.05604997	0.8129
report	1	1	1.20209324	0.2729
addresses	1	1	3.28251901	0.0700
free	1	1	95.5218652	<.0001*
business	1	1	22.4517501	<.0001*
email	1	1	1.08267106	0.2981
you	1	1	6.95181774	0.0084*
credit	1	1	12.8127096	0.0003*
your	1	1	21.3525061	<.0001*
font	1	1	1.92622793	0.1652
W_000	1	1	41.4959405	<.0001*
money	1	1	14.0255012	0.0002*
hp	1	1	122.442756	<.0001*
hpl	1	1	8.105804	0.0044*
george	1	1	244.834479	<.0001*
W_650	1	1	7.43763339	0.0064*

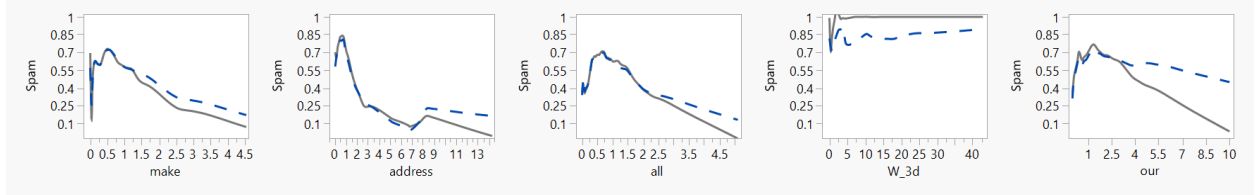
lab	1	1	9.39068957	0.0022*
labs	1	1	1.04769151	0.3060
telnet	1	1	0.246676	0.6194
W_857	1	1	0.58644706	0.4438
data	1	1	10.9999738	0.0009*
W_415	1	1	0.11452459	0.7351
W_85	1	1	8.00945184	0.0047*
technology	1	1	9.91500301	0.0016*
W_1999	1	1	0.09735046	0.7550
parts	1	1	3.71280334	0.0540
pm	1	1	7.09319583	0.0077*
direct	1	1	0.59187067	0.4417
cs	1	1	26.3995023	<.0001*
meeting	1	1	51.5123401	<.0001*
original	1	1	4.45190995	0.0349*
project	1	1	18.6795278	<.0001*
re:	1	1	47.8590513	<.0001*
edu	1	1	67.5471768	<.0001*
table	1	1	4.4269266	0.0354*
conference	1	1	16.6259729	<.0001*
C;	1	1	17.3685673	<.0001*
C(1	1	0.41144404	0.5212
C[1	1	0.92712056	0.3356
C!	1	1	32.1488028	<.0001*
C\$	1	1	86.6882651	<.0001*
C#	1	1	6.97683277	0.0083*
CAP_avg	1	1	0.30776473	0.5791
CAP_long	1	1	16.6221379	<.0001*
CAP_tot	1	1	21.5575672	<.0001*

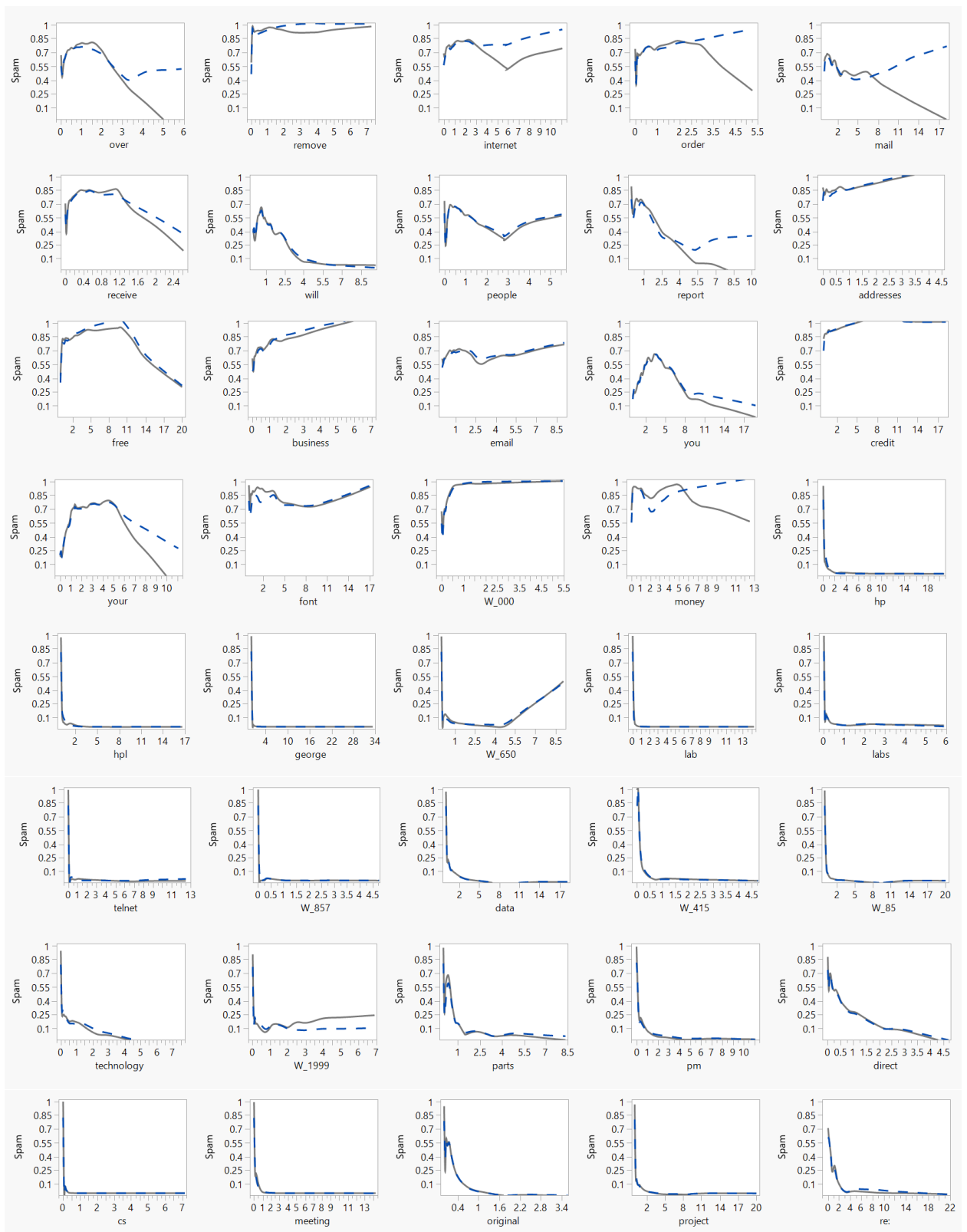
CONFUSION Matrix

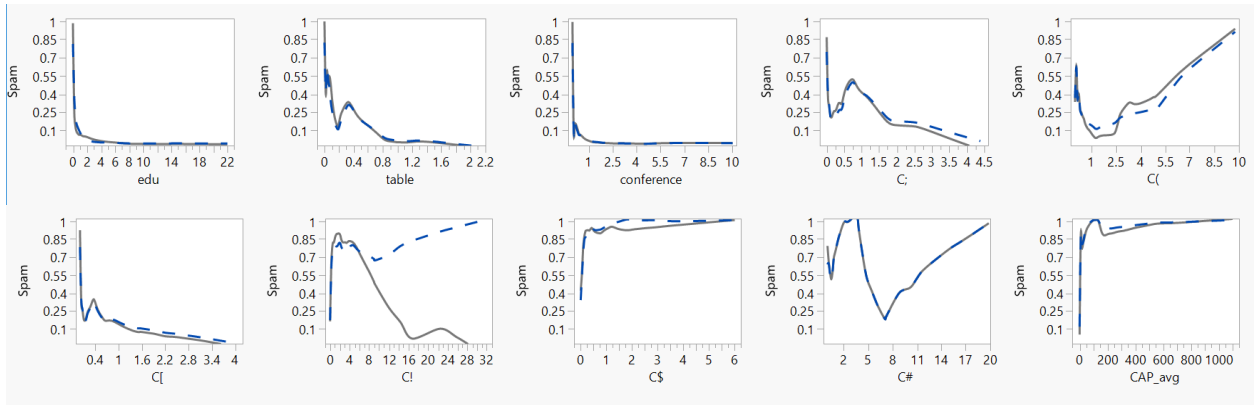
Actual \ Predicted	Count	
	1	0
Spam		
1	1610	203
0	122	2666

Sensitivity= 88.80

Specificity=92.92







Results after log transform.

Nominal Logistic Fit for Spam Effect Summary

Source	LogWorth	PValue
log(C!)	21.259	0.00000
log (our)	14.441	0.00000
C\$	12.847	0.00000
remove	9.547	0.00000
log (your)	9.382	0.00000
free	8.070	0.00000
hp	7.478	0.00000
george	7.006	0.00000
re:	6.169	0.00000
edu	5.636	0.00000
log (over)	5.057	0.00001
W_000	5.050	0.00001
CAP_tot	3.916	0.00012
internet	3.821	0.00015
business	3.427	0.00037
log (mail)	3.249	0.00056
meeting	3.020	0.00096
technology	2.957	0.00111
project	2.528	0.00296
log (order)	2.402	0.00397
money	2.401	0.00397
log (report)	2.339	0.00458
C;	2.268	0.00539
W_650	2.061	0.00869
W_85	2.027	0.00939
hpl	1.822	0.01506
conference	1.588	0.02585
will	1.505	0.03123
CAP_long	1.447	0.03573
C#	1.173	0.06720

Source	LogWorth																	PValue
CAP_avg	1.157																	0.06962
data	1.149																	0.07099
pm	1.127																	0.07473
credit	1.123																	0.07527
address	1.005																	0.09875
cs	0.991																	0.10216
lab	0.927																	0.11817
original	0.884																	0.13051
font	0.709																	0.19556
W_3d	0.706																	0.19656
make	0.697																	0.20111
addresses	0.632																	0.23338
table	0.596																	0.25365
C(0.581																	0.26228
email	0.556																	0.27786
parts	0.478																	0.33228
labs	0.447																	0.35750
direct	0.426																	0.37471
W_857	0.329																	0.46884
receive	0.318																	0.48071
people	0.315																	0.48452
C[0.233																	0.58500
you	0.227																	0.59227
W_415	0.190																	0.64572
telnet	0.101																	0.79286
W_1999	0.022																	0.95083
all	0.017																	0.96189

Converged in Gradient, 13 iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	2258.7604	57	4517.521	<.0001*
Full	826.3160			
Reduced	3085.0764			

RSquare (U)	0.7322
AICc	1770.14
BIC	2141.81
Observations (or Sum Wgts)	4601

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	4149	822.15716	1644.314

Source	DF	-LogLikelihood	ChiSquare
Saturated	4206	4.15888	Prob>ChiSq
Fitted	57	826.31605	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-2.1189372	0.1612309	172.72	<.0001*
make	-0.310679	0.2430236	1.63	0.2011
address	-0.1305942	0.0791037	2.73	0.0988
all	0.0058022	0.121445	0.00	0.9619
W_3d	2.10419066	1.6293828	1.67	0.1966
remove	2.03324215	0.3223539	39.78	<.0001*
internet	0.59688966	0.1575071	14.36	0.0002*
receive	-0.2082123	0.2952689	0.50	0.4807
will	-0.1668363	0.0774511	4.64	0.0312*
people	-0.1721486	0.2462591	0.49	0.4845
addresses	0.83369893	0.6995869	1.42	0.2334
free	0.81143367	0.1409255	33.15	<.0001*
business	0.812501	0.2283938	12.66	0.0004*
email	0.12971307	0.119537	1.18	0.2779
you	0.02119051	0.0395681	0.29	0.5923
credit	1.05868194	0.595151	3.16	0.0753
font	0.20510051	0.1584633	1.68	0.1956
W_000	2.07296734	0.4666618	19.73	<.0001*
money	0.45089815	0.1565406	8.30	0.0040*
hp	-1.744314	0.3158179	30.51	<.0001*
hpl	-1.1002774	0.4526284	5.91	0.0151*
george	-11.739381	2.2027467	28.40	<.0001*
W_650	0.50908177	0.193996	6.89	0.0087*
lab	-2.1157406	1.3540848	2.44	0.1182
labs	-0.2947379	0.3203202	0.85	0.3575
telnet	-0.1217606	0.4636823	0.07	0.7929
W_857	2.6302307	3.6310833	0.52	0.4688
data	-0.5403404	0.2992704	3.26	0.0710
W_415	-0.7699713	1.6748646	0.21	0.6457
W_85	-2.2310483	0.858953	6.75	0.0094*
technology	1.04888055	0.3215141	10.64	0.0011*
W_1999	-0.0109573	0.1776961	0.00	0.9508
parts	-0.413155	0.4261365	0.94	0.3323
pm	-0.6795931	0.3813373	3.18	0.0747
direct	-0.3483573	0.3924335	0.79	0.3747
cs	-46.402765	28.390379	2.67	0.1022
meeting	-2.8901554	0.8749041	10.91	0.0010*
original	-1.2909832	0.8537715	2.29	0.1305
project	-1.7748096	0.5972718	8.83	0.0030*
re:	-0.7707307	0.155145	24.68	<.0001*
edu	-1.2291967	0.2602014	22.32	<.0001*
table	-1.9014518	1.6657122	1.30	0.2537
conference	-3.7463104	1.6811065	4.97	0.0258*
C;	-1.2246388	0.4401188	7.74	0.0054*
C(-0.2911204	0.2596911	1.26	0.2623
C[-0.4586525	0.8398679	0.30	0.5850

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
C\$	5.2051259	0.7039509	54.67	<.0001*
C#	2.1251155	1.1610612	3.35	0.0672
CAP_avg	0.03567601	0.0196632	3.29	0.0696
CAP_long	0.00525452	0.0025021	4.41	0.0357*
CAP_tot	0.00087498	0.0002277	14.77	0.0001*
log (our)	1.56984393	0.1995386	61.90	<.0001*
log (over)	1.73160747	0.3895386	19.76	<.0001*
log (order)	1.31649139	0.4569925	8.30	0.0040*
log (mail)	0.66749115	0.1935454	11.89	0.0006*
log (report)	1.00165709	0.3533007	8.04	0.0046*
log (your)	0.84266799	0.1348618	39.04	<.0001*
log(C!)	2.37506185	0.2464184	92.90	<.0001*

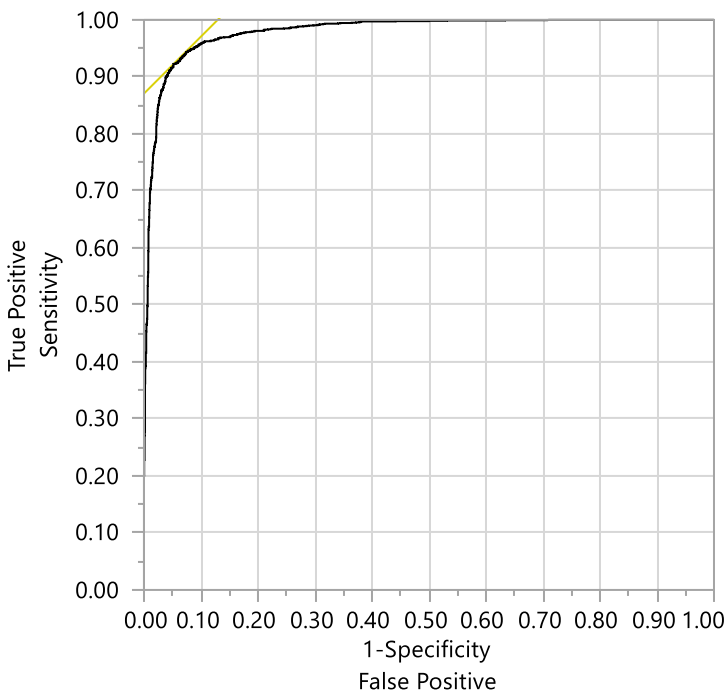
For log odds of 1/0

Effect Wald Tests

Source	Nparm	DF	Wald ChiSquare	Prob>ChiSq
make	1	1	1.63428198	0.2011
address	1	1	2.72554912	0.0988
all	1	1	0.00228258	0.9619
W_3d	1	1	1.6677231	0.1966
remove	1	1	39.7843594	<.0001*
internet	1	1	14.3611086	0.0002*
receive	1	1	0.49725304	0.4807
will	1	1	4.64008285	0.0312*
people	1	1	0.48867768	0.4845
addresses	1	1	1.42015319	0.2334
free	1	1	33.1533136	<.0001*
business	1	1	12.6554929	0.0004*
email	1	1	1.17750554	0.2779
you	1	1	0.28680822	0.5923
credit	1	1	3.16429272	0.0753
font	1	1	1.67523657	0.1956
W_000	1	1	19.7324238	<.0001*
money	1	1	8.29665122	0.0040*
hp	1	1	30.5053468	<.0001*
hpl	1	1	5.90909331	0.0151*
george	1	1	28.4028085	<.0001*
W_650	1	1	6.8863539	0.0087*
lab	1	1	2.4413673	0.1182
labs	1	1	0.84664861	0.3575
telnet	1	1	0.06895612	0.7929
W_857	1	1	0.52470506	0.4688
data	1	1	3.25992213	0.0710
W_415	1	1	0.21134389	0.6457
W_85	1	1	6.7465105	0.0094*
technology	1	1	10.6427031	0.0011*
W_1999	1	1	0.00380236	0.9508
parts	1	1	0.94000128	0.3323
pm	1	1	3.17599198	0.0747

Source	Nparm	DF	Wald ChiSquare	Prob>ChiSq
direct	1	1	0.78798416	0.3747
cs	1	1	2.67143939	0.1022
meeting	1	1	10.9124291	0.0010*
original	1	1	2.28642988	0.1305
project	1	1	8.82997591	0.0030*
re:	1	1	24.6791186	<.0001*
edu	1	1	22.3163598	<.0001*
table	1	1	1.30307895	0.2537
conference	1	1	4.96611837	0.0258*
C;	1	1	7.74241138	0.0054*
C(1	1	1.25669855	0.2623
C[1	1	0.29822606	0.5850
C\$	1	1	54.6736108	<.0001*
C#	1	1	3.35007726	0.0672
CAP_avg	1	1	3.29187774	0.0696
CAP_long	1	1	4.41004109	0.0357*
CAP_tot	1	1	14.7718899	0.0001*
log (our)	1	1	61.8954747	<.0001*
log (over)	1	1	19.7605001	<.0001*
log (order)	1	1	8.29885003	0.0040*
log (mail)	1	1	11.8939234	0.0006*
log (report)	1	1	8.03802034	0.0046*
log (your)	1	1	39.0422659	<.0001*
log(C!)	1	1	92.8974317	<.0001*

Receiver Operating Characteristic



Using Spam='1' to be the positive level

AUC
0.98011

Confusion Matrix

Training

Actual	Predicted Count	
Spam	1	0
1	1643	170
0	118	2670