ANTITRUST Notice

The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Variable Interactions in GLMs What Can Be Done?

Prof. Paul Beinat
Centre For Quantum Computation and Intelligent Systems
University of Technology Sydney
Director, NeuronWorks
Research and Development, EagleEye Analytics

Agenda

- Review of Current Methods
- Compounds to Ordinals
- Results
- Implications

Current Methods

- Constructing Variable Interactions
- Exhaustive Search
- Tree Based

Exhaustive Search

- Two Variables
  - Categoricals
    - Vehicle Body and Occupation
    - Combinations of all categorical values
      - Potentially many compound values - fragmentation
    - Heuristic Combinations
      - Limits combinations
      - Need heuristic theory – not easy
        - Heuristic = Analyst’s Bias
    - Signal Driven
      - Only generate potentially significant values
      - Fits noise

Exhaustive Search

- Two Ordinals
  - Driver Age and Horsepower
  - Combinations of bucketed values
    - Combined values depends on bucket numbers
    - Deriving buckets for use in combinations
      - Different to univariate buckets
  - Heuristic
    - Theory?
  - Signal driven
    - Noise?
  - Related to oblique splits

Exhaustive Search

- Categorical and Ordinal
  - Vehicle Body and Driver Age
  - Combinations of body and buckets of age
    - Potentially many values
  - Heuristic
  - Signal driven

Exhaustive Search

- Two variables
  - Many potential compound values
  - Strategies to limit value explosion
  - Many variables lead to
    - Many, many two variable interactions
  - Trivial example
- Three Variables?
- Four Variables?
- Come back next year!

Exhaustive Search

- Too Many Experiments
  - Well known phenomenon
  - Requires very careful use of validation data

- Not many compound variables derived this way make it into models
  - Too fragmented
  - Signal too weak
    - Overwhelmed by main effects variables

Tree Based

- Use Tree Induction
  - To identify leaf nodes of interacting variables
  - Potentially arbitrary complexity
- Implement
  - Whole tree as compound variable
  - Each two variable interaction as new variable
    - Almost none of these make it into model

Tree Induction

- Induction of Trees
  - Almost all research on classification
    - Little on regression
    - Nodes = piecewise constant function
      - Too few regression estimates
  - Implicit symmetric error distribution
  - Vectorized data
    - Averages dependent variable
  - Poor performance on insurance data
    - Asymmetric, noisy data

Compounds to Ordinals

- Ensembles
  - Combinations of weak learners
  - Bagging – Random Forests
    - Estimate is average of ensemble
    - Each member (tree) trained on bootstrap sample
  - Boosting – AdaBoost, TreeNet
    - Estimate is sum of ensemble
    - Each member trained on data minus previous cumulative model, or re-weighted data
  - Ensemble better than any learner member

Compounds to Ordinals

- Ensembles can overfit training data
  - Bootstrap samples contain similar data
  - Variables are often correlated
  - Must decorrelate ensemble members
  - Defence is randomization
    - Introduce randomization in tree
    - Stupid decisions improve performance

Compounds to Ordinals

- Random Forests and TreeNet
  - Use CART
  - Most research on classification
  - Models signal as interactions between variables
- Random Forests
  - Performs variance reduction
- TreeNet
  - Performs better on comparable data sets
  - Additive
    - Can produce negative estimates for insurance data

Compounds to Ordinals

- A new ensemble
- Re-derive Boosting
  - Multiplicative
    - Not additive
    - No negative estimates
  - Not additive in log space
    - Difficulty with premiums and claims

Compounds to Ordinals

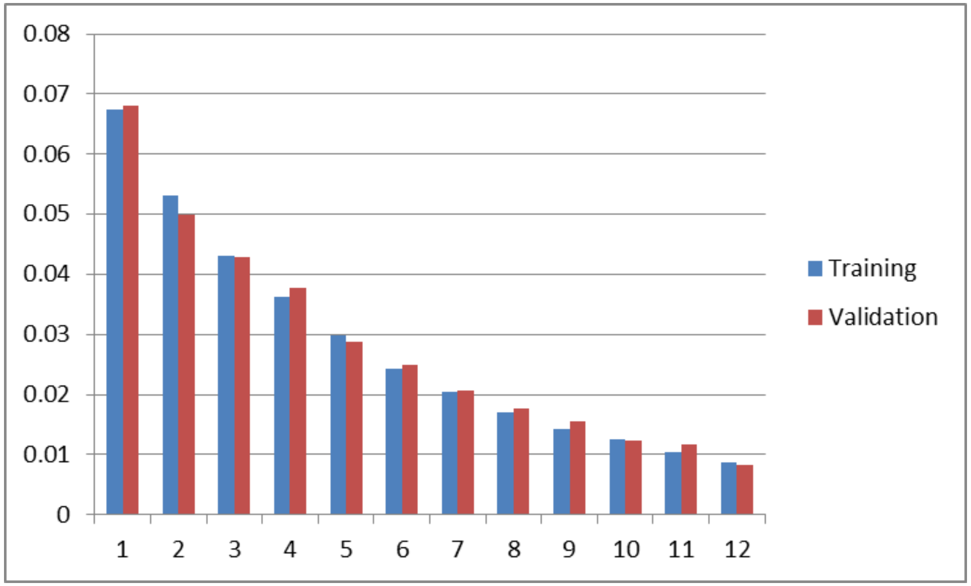
- Learner
  - Insurance specific induction
    - Exposure based observations
    - Premiums and claims
      - Loss ratio analysis
    - Multiple claims per exposure
    - Asymmetric errors
  - Strong learner – weakly applied
  - Output – score in 1 to 1000 range
    - Ensemble members combined into ordinal variable
  - Dimensionality reduction

Results

- Auto collision Book 1
- Derive claim frequency score
- Implement scores into GLMs
  - Forward stepwise
  - Measure on validation at policy level
- Training and Validation
  - 70% to 30%
  - At Random

Results 1 – Claim Frequency

Score Tier	Training Frequency	Validation Frequency
1-49	0.067466	0.068162
50 - 100	0.053092	0.049893
101 - 149	0.043059	0.042971
150 - 249	0.036161	0.037733
250 - 349	0.029762	0.028761
350 - 499	0.024245	0.025017
500 - 649	0.020408	0.02072
650 - 749	0.016974	0.017777
750 - 849	0.014281	0.01563
850 - 899	0.012504	0.01229
900 - 949	0.010371	0.011658
950 - 1000	0.008805	0.008288



Spread	0.01 - 0.07
Lift	7.824
Standard Deviation - Training	0.015
Standard Deviation - Validation	0.014
Correlation	0.998
Correlation - Exposure Weighted	0.997
F Statistic	864.265

Results 1 Claim Frequency

Variable	Influence
DriverAge	11.60%
Duration	11.60%
VehicleAge	9.07%
ClaimFreeYrs	9.07%
Occupation	6.33%
Channel	5.91%
NCD_Protection	5.27%
SecondCarDisc	5.06%
Excess	4.64%
Gender	3.80%
PPP	3.16%
Numdrivs	2.95%
MPGROUP	2.74%
Manufact	2.32%
Mileage_band	2.32%
Driving_Option	2.11%
Area	2.11%
Veh_value	1.90%
PenaltyPts	1.90%
Ccband	1.69%
disclm_free_yr	1.69%
Clsofuse	0.84%
CustomerDiscount	0.84%
Ncdlast	0.63%
DoorPlan	0.42%

Variable	Variable	Influence
DriverAge	ClaimFreeYrs	19.83%
DriverAge	Duration	19.83%
ageofveh	Duration	19.83%
DriverAge	VehicleAge	17.24%
Duration	Occupation	16.38%
Channel	Duration	15.52%
ClaimFreeYrs	Duration	15.52%
Excess	ClaimFreeYrs	14.66%
Duration	NCD_Protection	13.79%
DriverAge	Occupation	12.93%
Duration	SecondCarDisc	12.93%
VehicleAge	Channel	12.07%
VehicleAge	ClaimFreeYrs	11.21%
DriverAge	Gender	11.21%
NCD_Protection	SecondCarDisc	11.21%
Excess	DriverAge	10.34%
ClaimFreeYrs	Gender	10.34%
ClaimFreeYrs	Manufact	9.48%
ClaimFreeYrs	NCD_Protection	9.48%
VehicleAge	Occupation	9.48%
Channel	Occupation	9.48%
VehicleAge	Numdrivs	8.62%
VehicleAge	NCD_Protect	8.62%
Area	MPGROUP	8.62%
Gender	Manufact	8.62%

	able	Influence
	eage_band	7.76%
	ond_car_disc	7.76%
	altyPts	6.90%
		6.90%
	ndrivs	6.90%
	)_Protect	6.90%
	up	6.90%
	ndrivs	6.90%
	ond_car_disc	6.90%
	eage_band	6.90%
	der	6.90%
	)_Protect	6.90%
	nnel	6.90%
	ation	6.90%
	ing_Option	6.03%
	nufact	6.03%
	ndrivs	6.03%
		6.03%
	altyPts	5.17%
	up	5.17%
	up	5.17%
	GROUP	5.17%
	der	5.17%
	a	5.17%
AD_Excess	Channel	4.31%

	Variable	Influence
	Area	4.31%
	Driving_Option	4.31%
	NCD_Protection	4.31%
	MPGROUP	4.31%
	NCD_Protection	4.31%
	SecondCarDisc	4.31%
	SecondCarDisc	4.31%
	SecondCarDisc	4.31%
	SecondCarDisc	4.31%
	Veh_value	4.31%
	PPP	4.31%
c	Veh_value	3.45%
	Veh_value	3.45%
unt	SecondCarDisc	3.45%
	Numdrivs	3.45%
	Numdrivs	3.45%
	Occupation	3.45%
	PenaltyPts	3.45%
	PenaltyPts	3.45%
	PenaltyPts	3.45%
	PenaltyPts	3.45%
	PPP	3.45%
	PPP	3.45%
	Mileage_band	3.45%
ageofveh	Manufact	3.45%

Results 1 – Frequency GLM

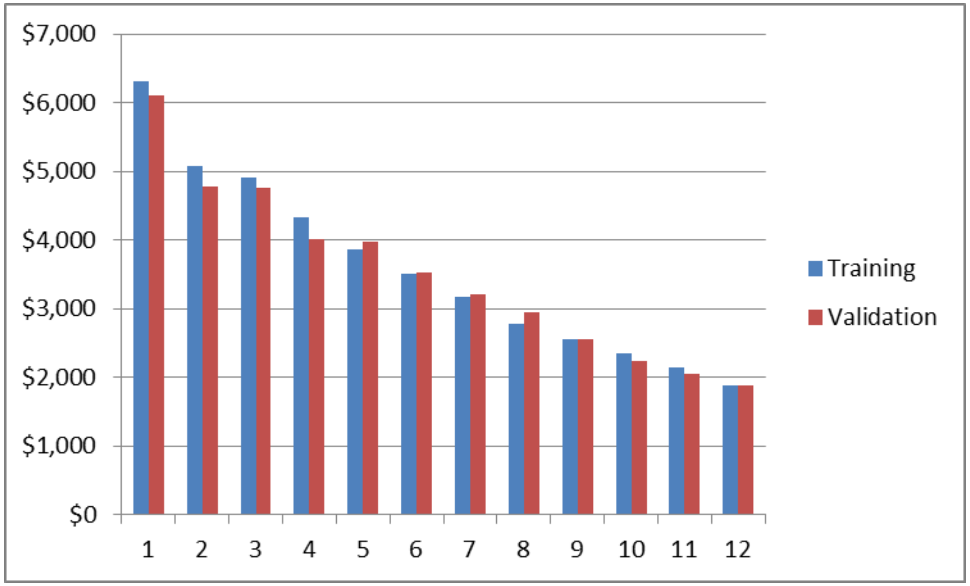
Without Score					With Score		
Iteration	Variable(s) Added	Deviance	Gini		Variable(s) Added	Deviance	Gini
1	NULL MODEL	109153.6	0		NULL MODEL	109153.6	0
2	NCD_Protection	108591.2	0.096974		FreqScore	106222.7	0.269983
3	disclm_free_yr	108059.1	0.145238		Licences	106092.2	0.275373
4	VehicleAge	107746.8	0.193252		PPP	106055.1	0.276167
5	Licences	107458.3	0.210782		NCD_Protection	106036.5	0.277947
6	PPP	107225.1	0.221158		disclm_free_yr	106023.3	0.279406
7	ClaimFreeYrs	107118.4	0.23429		VehicleAge	106017.2	0.281103
8	DriverAge	107051.4	0.239921		Ph_PenaltyPts	105981	0.282346
9	Ccband	106994.5	0.245086				
10	Driving_Option	106903.2	0.248889				
11	PenaltyPts	106822	0.252159				
23	DoorPlan	106606.3	0.265222				
24	Clsofuse	106604.3	0.265519				
25	CustomerDiscount	106604	0.265722				
26	AdditionalDrivers	106607.7	0.265944				
27	Mileage_band	106612	0.266103				
28	Ncdlast	106618.3	0.266489				
29	LoyaltyDiscount	106614.9	0.266652				
30	MPGROUP	106617.9	0.266717				
31	Auto_Manual	106618.8	0.266768				
32	Area	106619.8	0.266805				
33	Alarm_Immob	106620.6	0.266838				
34	Fuel_type	106620.2	0.266837				
35	Numdrivs	106620.2	0.266837				

Results

- Auto collision Book 1
- Derive severity score
- Implement scores into GLMs
  - Forward stepwise
  - Measure on validation at policy level
- Training and Validation
  - 70% to 30%
  - At Random

Results 1 - Claim Severity

Score Tier	Training Severity	Validation Severity
1-49	$6,313	$6,114
50 - 98	$5,085	$4,783
99 - 149	$4,907	$4,754
150 - 249	$4,334	$4,016
250 - 349	$3,872	$3,979
350 - 499	$3,502	$3,532
500 - 650	$3,171	$3,215
651 - 750	$2,787	$2,957
751 - 848	$2,557	$2,551
849 - 900	$2,356	$2,243
901 - 949	$2,149	$2,057
950 - 1000	$1,888	$1,883



Spread	1887 - 6250
Lift	3.312377
Standard Deviation - Training	1064.846
Standard Deviation - Validation	1004.103
Correlation	0.994996
Correlation - Exposure Weighted	0.99213
F Statistic	310.4135

Results 1 – Claim Severity

Variable	Influence
VehicleAge	14.94%
Ccband	12.33%
NCD_Protect	11.60%
Veh_value	10.76%
DriverAge	10.66%
Area	7.73%
Duration	6.06%
Excess	4.39%
Fuel_type	3.34%
Area1	2.30%
clm_free_yr	2.09%
MPGROUP	1.99%
Bodytype	1.67%
Gender	1.46%
PPP	1.46%
CustomerDiscount	1.25%
Ncdlast	1.15%
Manufact	0.84%
Numdrivs	0.73%
Auto_Manual	0.73%
PenaltyPts	0.42%
Driving_Option	0.42%
SecondCarDisc	0.31%
Clsofuse	0.31%
DoorPlan	0.31%
LoyaltyDiscount	0.31%
disclm_free_yr	0.21%
Mileage_band	0.21%

Variable	Variable	Influence
VehicleAge	Ccband	47.59%
VehicleAge	NCD_Protect	46.52%
DriverAge	VehicleAge	45.45%
Ccband	NCD_Protect	41.71%
VehicleAge	Veh_value	35.83%
DriverAge	Ccband	33.16%
VehicleAge	Area	31.55%
DriverAge	NCD_Protect	31.55%
NCD_Protect	Veh_value	31.02%
Ccband	Veh_value	28.34%
Area	Ccband	28.34%
DriverAge	Area	26.74%
DriverAge	Veh_value	26.74%
Area	NCD_Protect	25.13%
Duration	NCD_Protect	23.53%
DriverAge	Duration	21.93%
VehicleAge	Duration	20.86%
Ccband	Duration	20.32%
Duration	Veh_value	20.32%
Excess	Veh_value	17.65%
Excess	VehicleAge	17.11%
Area	Veh_value	16.58%
Fuel_type	Veh_value	15.51%
Area	Duration	13.90%
VehicleAge	Fuel_type	13.90%
Excess	NCD_Protect	13.90%
Area	Fuel_type	12.30%
DriverAge	Fuel_type	11.23%

ariable	Variable	Influence
D_Excess	Ccband	11.23%
uel_type	NCD_Protect	11.23%
geofDriver	Area1	10.70%
rea	Area1	10.16%
D_Excess	DriverAge	9.63%
geofveh	clm_free_yr	9.63%
cband	Fuel_type	9.63%
geofveh	Area1	8.56%
geofDriver	MPGROUP	8.56%
odytype	Veh_value	8.56%
rea	clm_free_yr	8.02%
cband	clm_free_yr	8.02%
D_Excess	Duration	7.49%
geofveh	MPGROUP	7.49%
odytype	MPGROUP	7.49%
lm_free_yr	NCD_Protect	7.49%
IPGROUP	Veh_value	7.49%
IPGROUP	NCD_Protect	6.95%
rea1	NCD_Protect	6.42%
geofDriver	Gender	6.42%
geofveh	Gender	6.42%
rea1	Ccband	6.42%
ICD_Protect	PPP	6.42%
geofDriver	clm_free_yr	5.88%
geofDriver	Bodytype	5.88%
geofveh	Ncdlast	5.88%
rea	Ncdlast	5.88%
Ccband	Ncdlast	5.88%

	Variable	Influence
	PPP	5.88%
	PPP	5.35%
	PPP	5.35%
	Gender	5.35%
	Fuel_type	5.35%
	Fuel_type	4.81%
	Duration	4.81%
	Bodytype	4.81%
	CustomerDiscount	4.81%
	NCD_Protect	4.81%
	MPGROUP	4.81%
	Ncdlast	4.81%
	Veh_value	4.81%
	Veh_value	4.28%
	Veh_value	4.28%
	NCD_Protect	4.28%
	NCD_Protect	4.28%
	NCD_Protect	4.28%
	Gender	4.28%
	Gender	4.28%
	Manufact	4.28%
	Manufact	4.28%
	Manufact	3.74%
	MPGROUP	3.74%
	Numdrivs	3.74%
	Numdrivs	3.74%
Ccband	Numdrivs	3.74%
NCD_Protect	Numdrivs	3.74%

Results 1 - Severity GLM

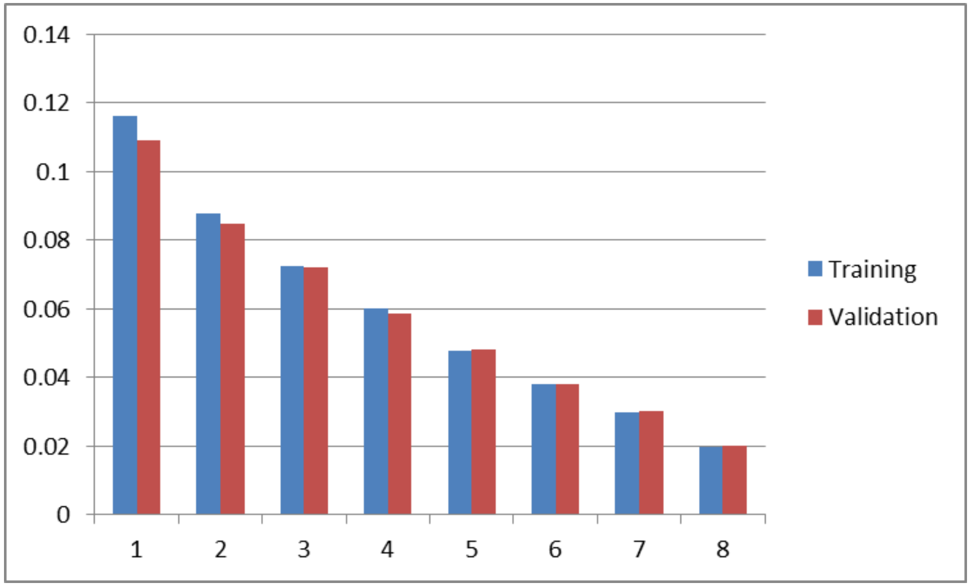
Without Scores					With Scores		
Iteration	Variable(s) Added	Deviance	Gini		Variable(s) Added	Deviance	Gini
1	NULL MODEL	7984.89	0		NULL MODEL	7984.89	0
2	VehicleAge	7647.402	0.077802		SevScore	6626.657	0.184246
3	MPGROUP	7458.309	0.110432		PPP	6613.677	0.186657
4	NCD_Protection	7300.899	0.128881		VehicleAge	6621.574	0.187986
5	Area	7226.274	0.138331		MPGROUP	6620.278	0.188379
6	Excess	7173.22	0.143735		DoorPlan	6609.287	0.188942
7	PPP	7141.864	0.147352		Auto_Manual	6609.934	0.189143
8	DriverAge	7098.941	0.151096		Ccband	6611.546	0.189555
9	Veh_value	7056.018	0.153923		AdditionalDrivers	6604.292	0.189846
23	Licences	6942.775	0.166879				
24	Auto_Manual	6941.092	0.16715				
25	Ncdlast	6941.542	0.167273				
26	CustomerDiscount	6938.596	0.167333				
27	SecondCarDisc	6937.987	0.167379				
28	ClaimFreeYrs	6939.362	0.167342				
29	disclm_free_yr	6940.159	0.167469				
30	Bodytype	6939.596	0.16763				
31	Ph_PenaltyPts	6940.794	0.167648				
32	Mileage_band	6939.733	0.167635				
33	PenaltyPts	6939.968	0.167642				
34	LoyaltyDiscount	6940.182	0.167634				
35	Numdrivs	6940.182	0.167634				

Results 2

- Auto collision Book 2
- Derive claim frequency score
- Derive frequency residual score from insurer GLM
- Implement scores into GLMs
  - Forward stepwise
  - Measure on validation at policy level
- Training and Validation
  - 70% to 30%
  - At Random

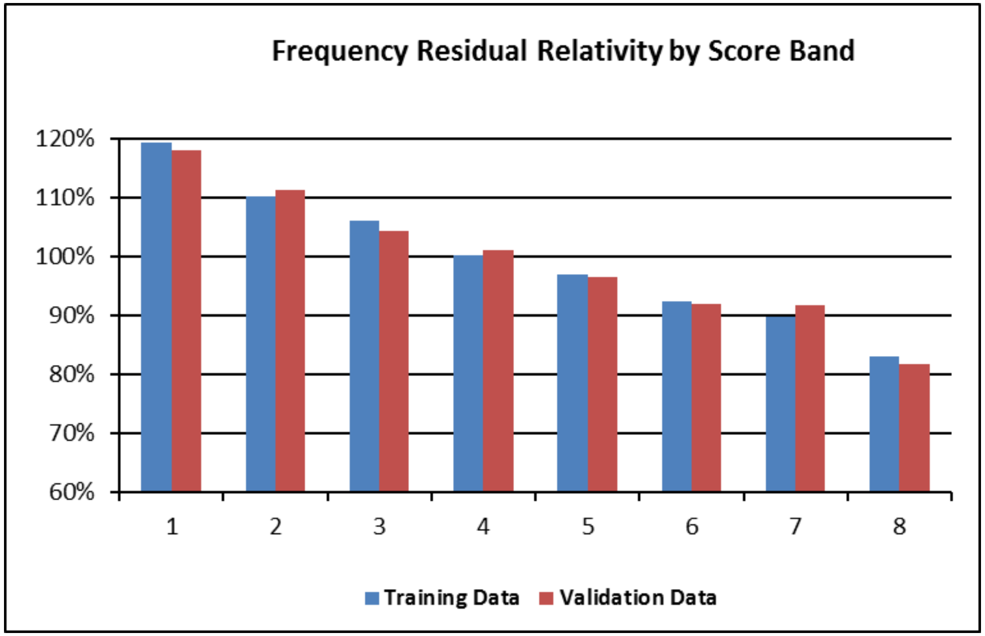
Results 2 – Claim Frequency

Score Tier	Training Frequency	Validation Frequency
1-49	0.116019	0.109042
50 - 149	0.087702	0.084814
150 - 299	0.072466	0.071968
300 - 499	0.060105	0.058669
500 - 699	0.047751	0.048239
700 - 849	0.038143	0.038181
850 - 949	0.029817	0.030271
950 - 1000	0.01982	0.019964



Spread	0.02 - 0.11
Lift	5.735187
Standard Deviation - Training	0.022747
Standard Deviation - Validation	0.021281
Correlation	0.99929
Correlation - Exposure Weighted	0.999147
F Statistic	1422.72

Results 2 – Claim Frequency Residual



Spread	0.82 - 1.19
Lift	1.440
Standard Deviation - Training	0.083
Standard Deviation - Validation	0.083
Correlation	0.994
Correlation - Exposure Weighted	0.991
F Statistic	349.226

Score Tier	Exposure	Frequency	GLM Estimate	Frequency Residual
1-49	103,583	6.16%	5.18%	1.188
50 - 149	210,676	6.63%	6.01%	1.104
150 - 299	316,386	6.56%	6.23%	1.054
300 - 499	421,708	6.43%	6.41%	1.004
500 - 699	421,788	5.82%	6.02%	0.967
700 - 849	315,876	4.70%	5.09%	0.922
850 - 949	210,731	4.02%	4.45%	0.903
950 - 1000	107,648	2.55%	3.10%	0.825
	2,108,398	*5.64%*	*5.64%*	*0.999*

Systematic and consistent frequency residual by score band

Results 2 - Forward GLM

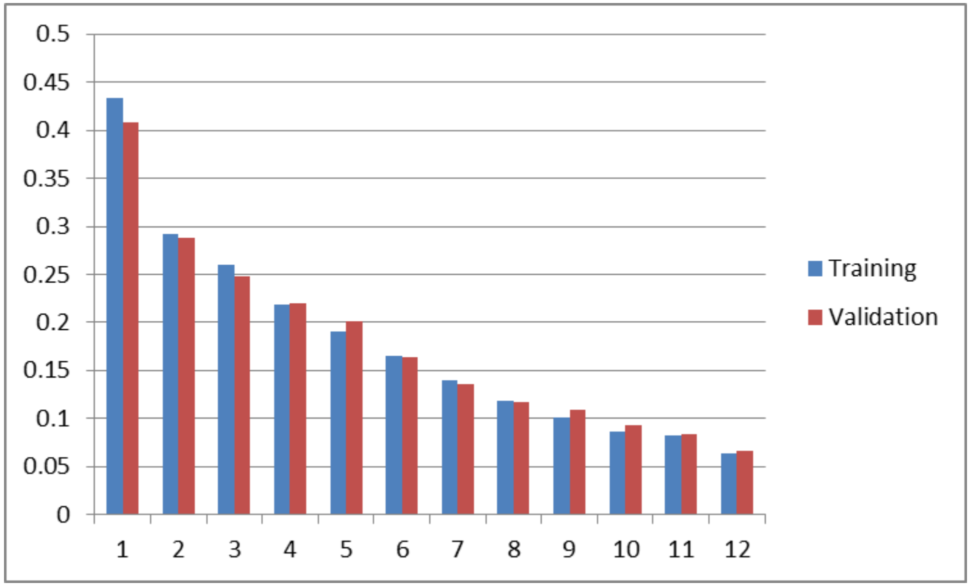
Iteration	Variable(s) Added	Deviance	Gini	Log Likelihood		Iteration	Variable(s) Added	Deviance	Gini	Log Likelihood
1	NULL MODEL	238291.7	0	-153679		1	NULL MODEL	238291.7	0	-153679
2	v9	237082.2	0.09669	-153075		2	FreqScore	233119.6	0.223941	-151093
3	d6	236286.1	0.136722	-152677		3	p2	232900	0.228215	-150984
4	g1	235545.5	0.15616	-152306		4	d6	232817.7	0.230518	-150942
5	p7	235165.6	0.168049	-152116		5	FreqResidualScore	232719.5	0.2323	-150893
6	p2	234852.2	0.17599	-151960						
7	p19	234507	0.183583	-151787						
8	p11_ord	234283.8	0.189785	-151675						
9	v10	234155	0.193973	-151611						
10	d13_ord	233979.6	0.196758	-151523						
31	p9	232837.6	0.221719	-150952						
32	v4	232843.3	0.221926	-150955						
33	d3	232851.2	0.222033	-150959						
34	p4	232850.6	0.222107	-150959						
35	v17	232841.8	0.222201	-150954						

Results 3

- Third party liability auto coverage
- Derive claim frequency score
- Implement scores into GLMs
  - Forward stepwise
  - Measure on validation at policy level
- Training and Validation
  - 70% to 30%
  - At Random

Results 3 - Claim Frequency

Score Tier	Training	Validation
1-49	0.432858	0.407639
50 - 99	0.292108	0.287672
100 - 149	0.259472	0.248512
150 - 249	0.21807	0.219564
250 - 349	0.190543	0.200586
350 - 499	0.164658	0.164284
500 - 649	0.140098	0.1362
650 - 749	0.118803	0.117711
750 - 849	0.100806	0.108623
850 - 899	0.086426	0.092492
900 - 949	0.082022	0.083861
950 - 1000	0.063426	0.06655



Spread	0.06 - 0.43
Lift	6.610288
Standard Deviation - Training	0.083656
Standard Deviation - Validation	0.078206
Correlation	0.998466
Correlation - Exposure Weighted	0.997427
F Statistic	660.3628

Results 3 – Forward GLM

No Frequency Score				With Frequency Score			
Iteration	Variable(s) Added	Deviance	Gini	Iteration	Variable(s) Added	Deviance	Gini
1	NULL MODEL	218925.1	0	1	NULL MODEL	218925.1	0
2	BonusMalus	215573.3	0.16932	2	FreqScore	210961.8	0.255272
3	payments per term	214228.4	0.202165	3	BonusMalus	210866	0.257335
4	Max Limit	213033	0.223193	4	Driver Age 2	210821.6	0.258301
5	Age of Vehicle	212639.2	0.230148	5	Driver Age	210790.1	0.258827
6	Driver Age	212232	0.236614	6	type of chassis	210777.2	0.259121
7	postal code	211967.3	0.240143	7	Age of Vehicle	210784.2	0.259273
8	fuel	211620.5	0.243306	8	Policy Discount	210780.6	0.259403
9	Driver Age 2	211432.2	0.245733	9	KW	210769.8	0.259565
10	HORSEPOWER	211360.6	0.246804	10	HORSEPOWER	210759.3	0.25969
11	Policy Discount	211325.5	0.247825	11	postal code	210743.9	0.259821
12	KW	211285.7	0.248275	12	Max Limit	210722.8	0.259897
13	type of chassis	211268.2	0.248708	13	fuel	210703.2	0.259916
14	sex	211259.9	0.248774	14	payments per term	210696	0.259931
15	Years	211259.9	0.248774	15	sex	210692.1	0.259949
16	Limit Amount	211259.9	0.248774	16	Years	210692.1	0.25995
17	type of chassis	211259.9	0.248774	17	Limit Amount	210692.1	0.25995
18	fuel 2	211259.9	0.248774	18	fuel 2	210692.1	0.25995

Implications

- Multiplicative boosted ensembles
  - Produce compound variables as scores
  - Scored compound variables are very powerful
  - Claim frequency, claim severity and loss ratio
  - Similar results found for other lines
- Why does this happen?
  - It is not an accident
    - The world really is compound and complex
  - Many compound variables combined into one framework
    - Avoids fragmentation – reduces dimensions