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.

Machine Learning

Insurance Applications

Panelists: Dr. Paul Beinat, Research and Development, EagleEye Analytics

Machine Learning

- Example Result
- Why do it?
- Types of problem classes
- Data issues
- Methods
- Testing
- Results

Example Result

Segment	Exposure	GLM Total Premium	Claims	Claim Count	Training LR
1	40,088	9,677,889	7,223,230	5,730	75%
8	26,642	8,770,620	7,454,508	4,717	85%
3	35,946	8,036,238	7,298,945	5,178	91%
4	20,954	6,699,637	6,353,455	3,664	95%
6	26,212	6,754,957	6,534,512	4,127	97%
0	29,558	7,868,872	8,109,686	5,018	103%
9	20,049	5,636,667	5,935,182	3,576	105%
2	33,043	10,830,010	11,614,780	6,287	107%
7	23,203	8,181,896	10,125,938	4,356	124%
5	30,163	7,419,663	9,590,068	5,081	129%

Why do Machine Learning?

- Database resources are largely going to waste
  - MIS, DSS are not machine learning
- Increasing rate of data collection
- Knowledge Based Systems
  - Domain Experts have limited knowledge
  - Knowledge Acquisition is a slow process

Machine Learning Myths

- Machine learning tools need no guidance.
- Machine learning requires no data analysis skill.
- Machine learning eliminates the need to understand your business and your data.
- Machine learning tools are “different” from statistics.

Types of Problems

- Classification
  - Fraud detection, loan approval, etc.
- Regression
  - Insurance rating, stock price prediction, etc.
- Clustering (Pattern Detection)
  - Customer clustering, time series, etc
- Mixtures

Data

- Data Types
- Reliability
  - Missing data
  - Skewed data
  - Noisy Data
- Variability
  - Data changes over time
- Sufficiency
  - Sample size
- Testing implications
  - Validation data

Classification Learning

- Supervised Learning, Inductive Learning
- Dataset contains examples of input attributes and their corresponding classification

How Does It Work?

- Examine every input attribute
  - Find the best split that can be made
- Select the best attribute, with the best split
- Build a branch for it and assign the appropriate subset of the data
- Determine if any branch is a leaf node
- Send that subset back to the start

Limitations

- No guaranteed way to find optimal solution
- Induction must compromise
  - Specificity
  - Simplicity
- OR
  - Accuracy
  - Description length

Example

Air Pressure	Ambience	Filtering	Air Flow	Type
988.6	Medium	Moderate	Light	A
989.3	Low	Mild	Light	B
992.3	Medium	Severe	Extreme	B
993.1	High	Severe	Fast	B
996.5	High	Mild	Light	A
996.9	Low	Average	Moderate	B
997.4	Medium	Moderate	Moderate	A
998.6	High	Severe	Fast	B

- If Ambience is low -> type is B
- If Ambience is medium and Filtering is moderate -> Type is A
- If Ambience is medium and Filtering is severe -> Type is B
- If Ambience is high and Filtering is mild -> Type is A
- if Ambience is high and Filtering is severe ->Type is B

Good and Bad of Induction

- Many input attributes
- Easy to understand results
- Relatively fast to run
- Relatively easy to use
- Several algorithms to choose from
  - IDS, C4.5, CART, O-Btree
- No non-linear effects
- No optimal way of partitioning numeric attributes
- Poor performance on noisy data
- No algorithm selection rules found
- Regression uncommon

Knn - K Nearest Neighbour

- Take k nearest instances to make estimate

- Issues
  - How many is k
  - How far away can neighbors be
  - CBR

How Many in k

- Static - determined at development time

- Dynamic
  - Statistically based
  - Heuristically based
  - Composites
    - With/without Gaussian biases
    - With/without directional imperatives

How Far Away for Neighbours

- Euclidian distance
  - Difficult with categorical data
  - Development of biases for axes
- Heuristic distance
  - Encodes domain knowledge
- Non-linear
  - Interaction between variables

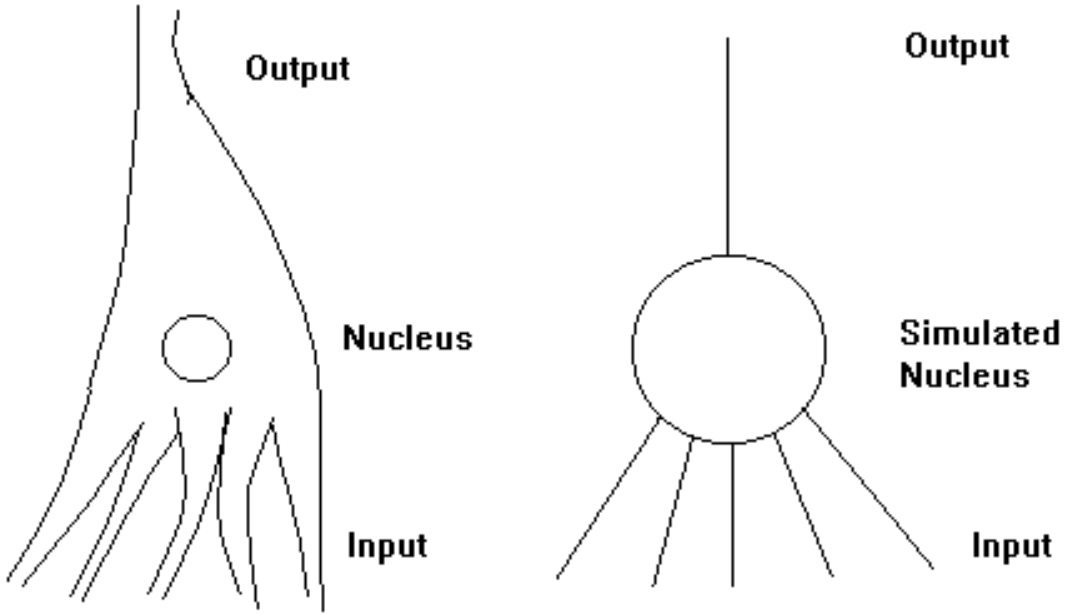
Classification

- Easy when all neighbors agree
- Problems
  - Sparsity of data locally
  - Neighbors not in agreement
    - Probabilistic democratic classification
    - Weighted probabilities base on distance
  - Edges - all neighbors on one side

Regression

- Central estimate based on neighbours
- Estimate biased by distance
- Certainty
  - Sample sufficiency
  - Distribution
- Edges
  - Incorporate multi-dimensional linear trends
- Care with skewed distributions

Neurons - Real v Artificial



Artificial Neural Networks

- Fundamentals
- Training
- Problems with NN's
- NN Variants

Fundamentals

- Each Neuron
  - A Number of Inputs
    - Each Input Has a Weight Associated With It
    - The Weight Moderates Sensitivity To That Input
  - One Output
    - The Output Can be Connected to Many Neurons
    - As Input
  - A Mathematical Function Inside
    - Controls Firing

Training

- Initialization
  - Weights Must be Randomized
- Learning Function
  - Back propagation
  - Weight Space and Error Function
- Data
  - Training Set
  - Test Set
  - OverFitting

Using Neural Nets

- Set learning parameters
  - Learning rate, momentum
  - Tolerance
- Train
- Test
- Iterate and tune

Problems?

- Local Minima!
  - Shaking
  - Genetic Algorithms
- OverFitting
  - Needs validation data to control
- Topology
  - Brittleness to Change (Sine Function NN)

More Problems

- Black Box
  - No Insight
  - No Expert Validation
- No Domain Knowledge
  - No Ability to Use the Obvious

Adaptive Networks

- Cascade Correlation
  - Start With Inputs Connected to Outputs - Train
  - Add One Hidden Node - Set Orthogonal to Existing Nodes - Retrain Other Nodes
  - Go Back to Step 2
  - Stop When Adding a Node Does Not Improve Fit
  - Can Produce Mapping with Abrupt Transitions

Optimal Brain Damage

- Start With Overly Populated NN – Train
- Select Hidden Layer Node with Least
- Sensitivity to Input Nodes
- Delete that Node - Re-Train
- Repeat
- Stop When Next Iteration Does Not Improve Fit

Radial Basis Functions

- Hidden Units Use Gaussian Activation
- Activation Governed by Euclidian Distance Between Weight and Input Vectors
- Hidden Units Represent Clusters in Data
- Gradient Descent Learning
- Input Attribute Numbers Cause Geometric Explosion of Hidden Units

Kohonen Self Organising Maps

- Unsupervised, Winner Take All Learning (No Ouput Values)
- Clustering Problem
- Units Arranged in 2 Dimensional Lattice
- Weight Vector Same Dimensionality as Input Vectors
- Randomize Weight Vectors Initially

SOM's

- For Each Input Vector
  - Determine Which Unit Wins
  - Change Its Weight Vector Towards Input Vector (by Learning Rate)
  - Change All Its Neighbors (Limited by Neighborhood Size) Towards Input Vector
- As Training Proceeds
  - Reduce Neighborhood Size
  - Reduce Learning Rate

Results

- Methodology
  - Use machine learning techniques to search the residuals of the GLMs
  - Medium size auto portfolio
  - Pure premiums set by consulting actuaries using GLM tools on 4 years of data
  - Pure premiums derived from ALL of the data
  - We divide data into training and validation
  - Search for difference between pure premiums and claims signal

Results

- Methodology
  - Use only policy attributes
  - One year for training, subsequent year for validation
  - Derive point estimates for segments of the portfolio using a minimum data requirement (number of claims)
  - Derive continuous estimates
  - Test accuracy of estimates

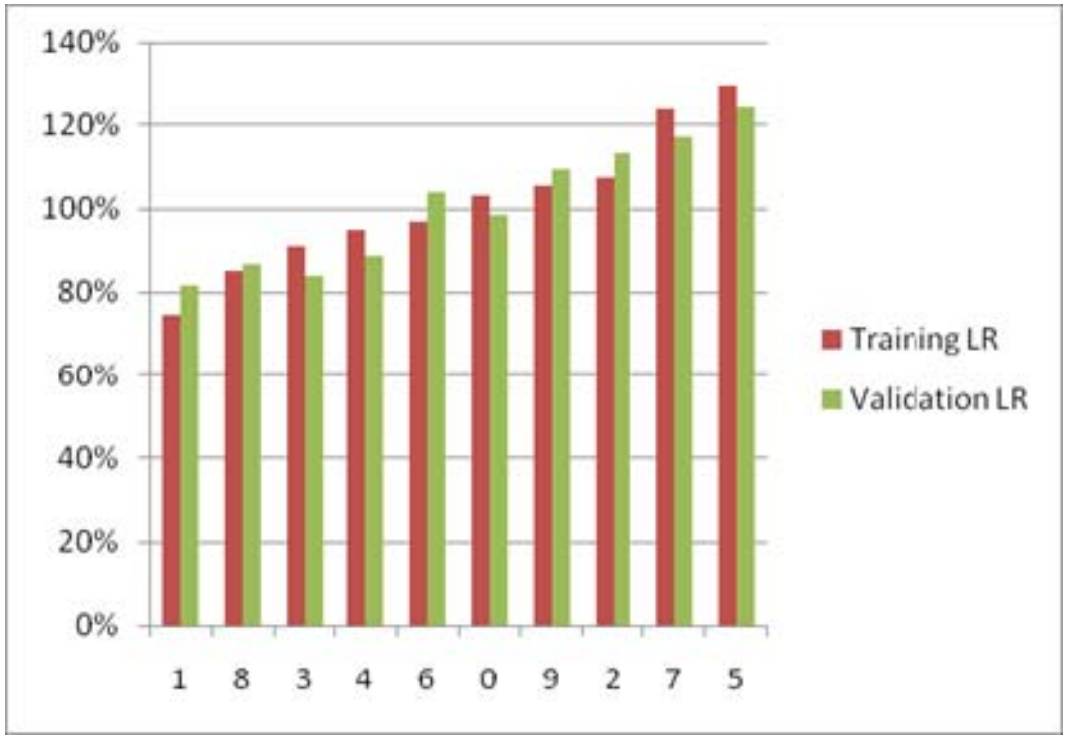
Consulting Actuaries A Training Results

Segment	Exposure	GLM Total Premium	Claims	Claim Count	Training LR
1	40,088	9,677,889	7,223,230	5,730	75%
8	26,642	8,770,620	7,454,508	4,717	85%
3	35,946	8,036,238	7,298,945	5,178	91%
4	20,954	6,699,637	6,353,455	3,664	95%
6	26,212	6,754,957	6,534,512	4,127	97%
0	29,558	7,868,872	8,109,686	5,018	103%
9	20,049	5,636,667	5,935,182	3,576	105%
2	33,043	10,830,010	11,614,780	6,287	107%
7	23,203	8,181,896	10,125,938	4,356	124%
5	30,163	7,419,663	9,590,068	5,081	129%

Consulting Actuaries A Validation Results

Segment	Exposure	GLM Total Premium	Claims	Claim Count	Validation LR
1	39,262	9,511,229	7,767,501	5,913	82%
8	20,083	6,415,686	5,565,564	3,784	87%
3	35,105	7,505,323	6,283,145	5,073	84%
4	15,379	4,749,230	4,195,864	2,822	88%
6	29,387	6,935,811	7,187,731	4,688	104%
0	33,141	8,311,156	8,171,977	5,761	98%
9	20,488	5,266,095	5,748,663	3,720	109%
2	34,729	10,911,435	12,336,791	6,768	113%
7	24,679	8,140,954	9,532,883	4,641	117%
5	25,717	5,925,355	7,358,740	4,570	124%

Consistency



How well does it do?

- Correlation 0.93
- Lift
  - Using exposure weighted standard deviation of loss ratios
  - Training 16.4%
  - Validation 14.5%
  - Minimal overfit
- Result
  - Consistent and large signal present in GLM residuals

How well does it fit?

	Deviance	Squared error	Chi squared error
GLM Premiums	34.15388	1463.838	5.47708
Estimated Premiums	15.02064	243.8955	0.946702

- Using validation data
- GLM estimates optimistic, validation used in training (trained on all 4 years of data)
- Derive estimated premiums, use relativities derived from training and applied to validation
- Fit much better than GLM premiums

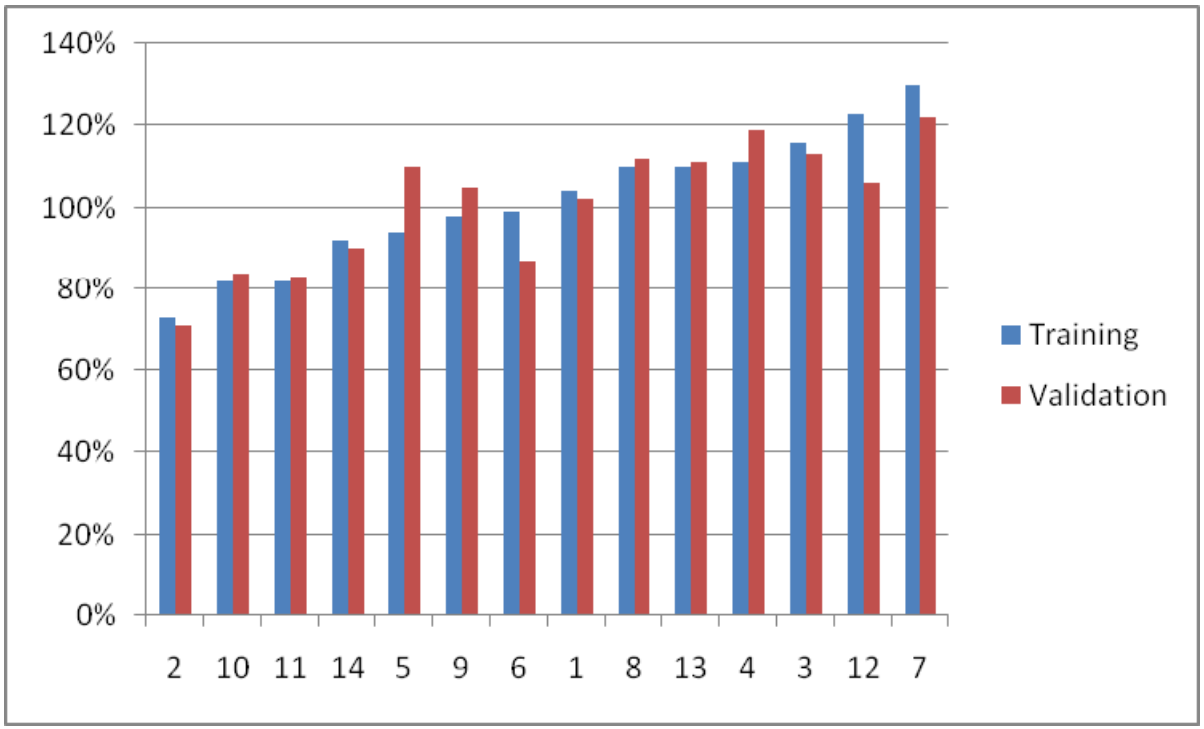
Consulting Actuaries B Training Results

Segment #	Exposure	GLM Total Premium	Claims	Claim Count	Training LR
2	35810	9334790	6828787	3937	73%
10	18123	4598518	3761181	2316	82%
11	14747	4841290	3965517	2130	82%
14	12735	5087922	4692275	2143	92%
5	16357	5754276	5410582	2451	94%
9	21869	6447323	6324888	3081	98%
6	13881	4716065	4646039	2164	99%
1	28572	8969315	9290752	3943	104%
8	21395	6571921	7257317	3216	110%
13	24882	6207643	6808248	3428	110%
4	17983	7213784	7982848	2865	111%
3	34028	7286684	8441279	3994	116%
12	14673	5198631	6382962	2786	123%
7	17577	3368306	4389912	2360	130%

Consulting Actuaries B Validation Results

Segment #	Exposure	GLM Total Premium	Claims	Claim Count	Validation LR
2	30881	8314942	5880703	3409	71%
10	21658	6049587	5054876	2829	84%
11	16037	5806587	4809694	2316	83%
14	16239	7151732	6406690	2847	90%
5	17261	6793896	7482004	2909	110%
9	23830	7897427	8271474	3734	105%
6	16552	6079428	5316744	2582	87%
1	28601	9996409	10161060	4141	102%
8	22232	7446225	8333203	3529	112%
13	23657	6443905	7159838	3344	111%
4	18130	8232309	9807524	3239	119%
3	31881	7379681	8362041	3860	113%
12	16120	6356732	6738014	3176	106%
7	16800	3478195	4253857	2114	122%

Consistency



How well does it do?

- Correlation 0.87
- Lift
  - Using exposure weighted standard deviation of loss ratios
  - Training 16.3%
  - Validation 15.2%
  - Minimal overfit
- Result
  - Consistent and large signal present in GLM residuals

How well does it fit?

	Deviance	Squared error	Chi squared error
GLM Premiums	41.60947	2241.3318	7.9576
Estimated Premiums	18.30203	737.3713	1.814276

- Using validation data
- GLM estimates optimistic, validation used in training (trained on all 4 years of data)
- Derive estimated premiums, use relativities derived from training and applied to validation
- Fit much better than GLM premiums

Continuous Estimates

- Estimates made of exposures based on a 0 to 1000 range
  - 0 is best loss ratio
  - 1000 is worst loss ratio
  - An insurance score – parallel to credit score

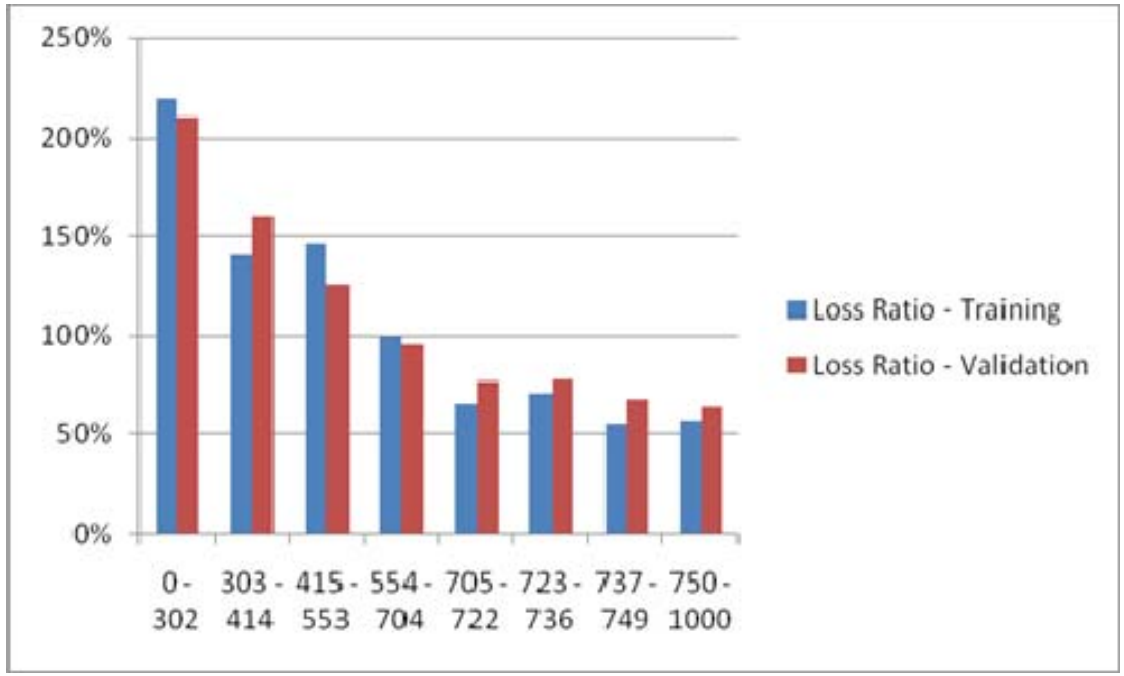
Training Results

Output Range	Exposure	Premium	Claims Cost	Claim Count	Loss Ratio
0 - 302	7324	1,802,287	3,949,415	1301	219%
303 - 414	14569	4,079,891	5,779,738	2440	142%
415 - 553	20790	5,756,714	8,394,295	3537	146%
554 - 704	173637	52,046,796	51,689,134	26266	99%
705 - 722	27797	7,020,762	4,640,641	3295	66%
723 - 736	19114	4,439,842	3,151,909	2100	71%
737 - 749	14143	3,092,726	1,724,326	1395	56%
750 - 1000	8743	1,707,370	969,907	769	57%

Validation Results

Output Range	Exposure	Premium	Claims Cost	Claim Count	Loss Ratio
0 - 302	9598	2,599,238	5,468,544	1901	210%
303 - 414	13563	3,804,456	6,090,286	2666	160%
415 - 553	18323	4,985,779	6,285,874	3142	126%
554 - 704	163770	46,299,639	44,418,782	24868	96%
705 - 722	28430	6,788,571	5,254,515	3429	77%
723 - 736	20059	4,400,802	3,421,952	2305	78%
737 - 749	14880	3,081,390	2,109,845	1569	68%
750 - 1000	9644	1,788,207	1,156,329	899	65%

Consistency



How well does it do?

- Correlation 0.98
- Lift
  - Using exposure weighted standard deviation of loss ratios
  - Training 30.4%
  - Validation 28.9%
  - Minimal overfit
- Result
  - Consistent and large signal present in GLM residuals

How well does it fit?

	Deviance	Squared Error	Chi Squared Error
GLMs	44.75	5722	21.74
Output Ranges	18.23	528	1.99

- Using validation data
- GLM estimates optimistic, validation used in training (trained on all the data)
- Derive estimated premiums, use relativities derived from training and applied to validation
- Fit much better than GLM premiums

Summary

- Point estimates (a piecewise constant function) derived from training data fit the validation data much better than the GLM
  - Improvement in fit is very significant
- Regardless of who fits the GLM
- Continuous estimates (scores) also fit validation data better than the GLM
  - Have greater lift
  - More lift and better fit than credit scores

Conclusion

- GLMs
  - Overfit their beta values (see model validation)
  - Underfit the signal in the data
    - Locally acting highly compound variables
    - Surely consulting actuaries didn't mean to underfit the signal to this extent
- Machine learning methods can be a valuable supplement to GLMs
  - Can find extra lift
  - Can improve fit of pure premiums

Implication

- Current GLM based cost models are inaccurate
- Extrapolation from an inaccurate cost model is unreliable
- Price optimization approach produces unpredictable results
  - Could reduce price on current loss makers in order to increase their market share!