18TH EAST ASIAN ACTUARIAL CONFERENCE
12-15 October 2014
Taipei International Convention Center in Taipei Taiwan

Predictive Modelling: GLM vs Machine-Learning

Guanjun Jiang
Principal & Consulting Actuary
Milliman Limited

Agenda

- Introduction of Predictive Modelling
- Generalised Linear Model (GLM)
- Machine Learning (Eagle Eye Analytic)
- Case Study
- Summary

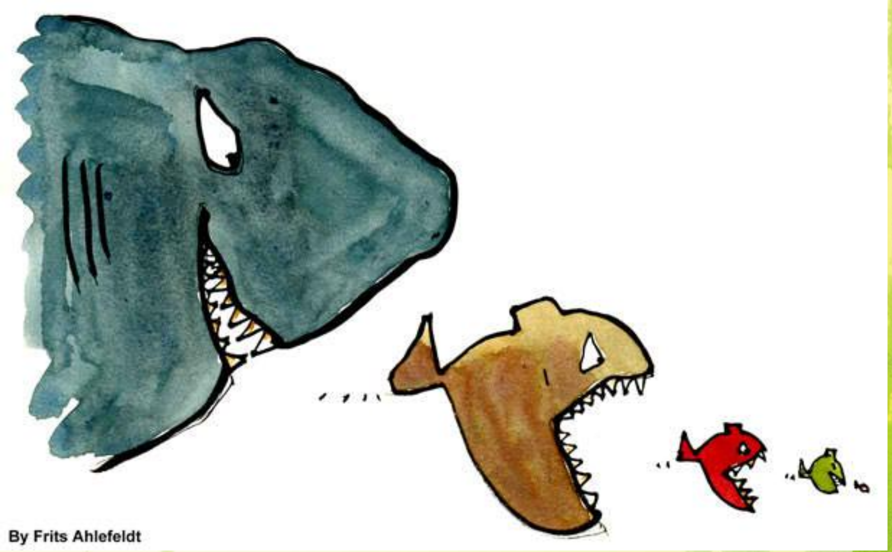
PREDICTIVE MODELLING

—— PREDICTIVE MODELLING IS THE PROCESS BY WHICH A MODEL IS CREATED OR CHOSEN TO TRY TO BEST PREDICT THE PROBABILITY OF AN OUTCOME.

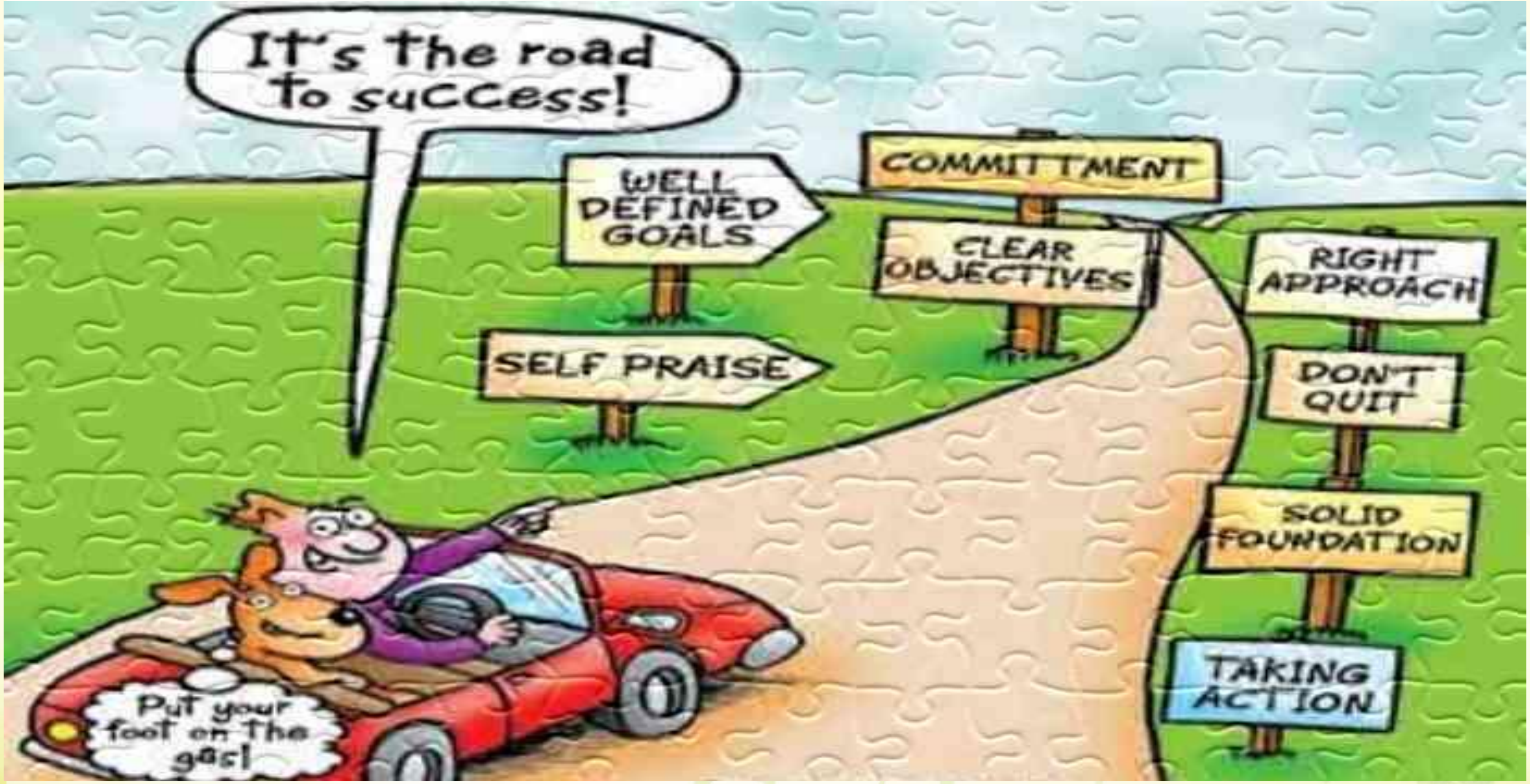
Some Insurers……

Are happy doing what they have always done…….

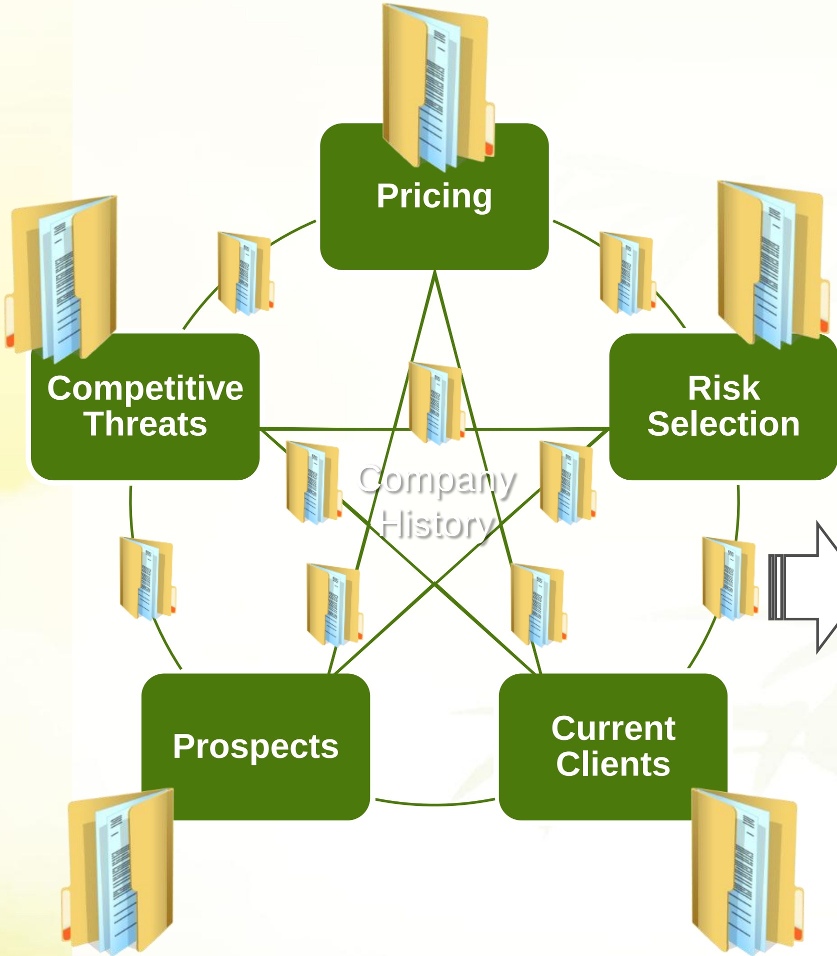
Others are happy being better and bigger than some……..



But Today Predictive Analytics Can Give You A Clear Vision of Where You Are and Light The Path Ahead.



Key Areas of Business Interaction



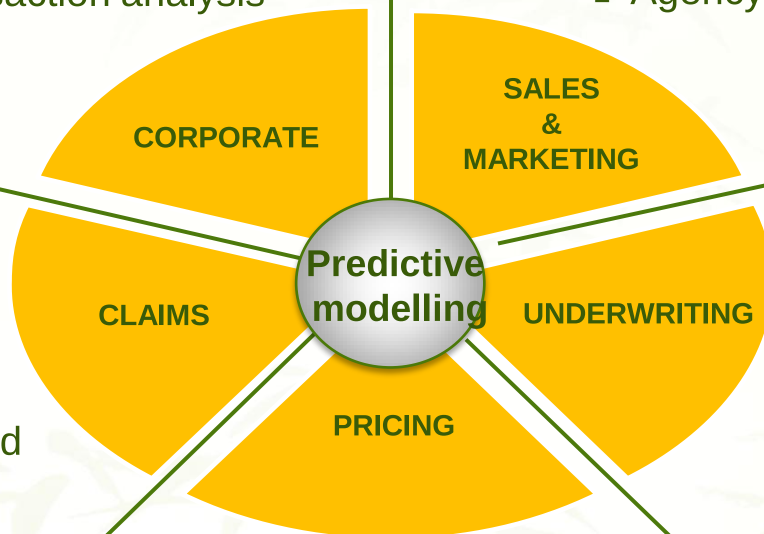
Knowledge gained results In competitive advantage through:

- Greater client satisfaction and retention
- Better risk selection
- Granular, targeted pricing
- More effective marketing

Usages of Predictive Modelling in Insurance

- Underwriting cycle management
- Profitability Analysis
- Reinsurance optimization
- M&A post-transaction analysis

- Target marketing
- New business acquisition
- Retention management
- Agency management



- Claims routing and prioritization
- High risk identification
- Loss control
- Reserve projection and estimation

- Risk Selection/avoidance
- Objectivity/consistency
- Risk scoring
- Loss control

- Accuracy & adequacy
- Competitiveness
- Adverse Selection
- Customer view
- Efficiency

GENERALISED LINEAR MODEL

Brief Introduction of GLM

- Basic Structure:

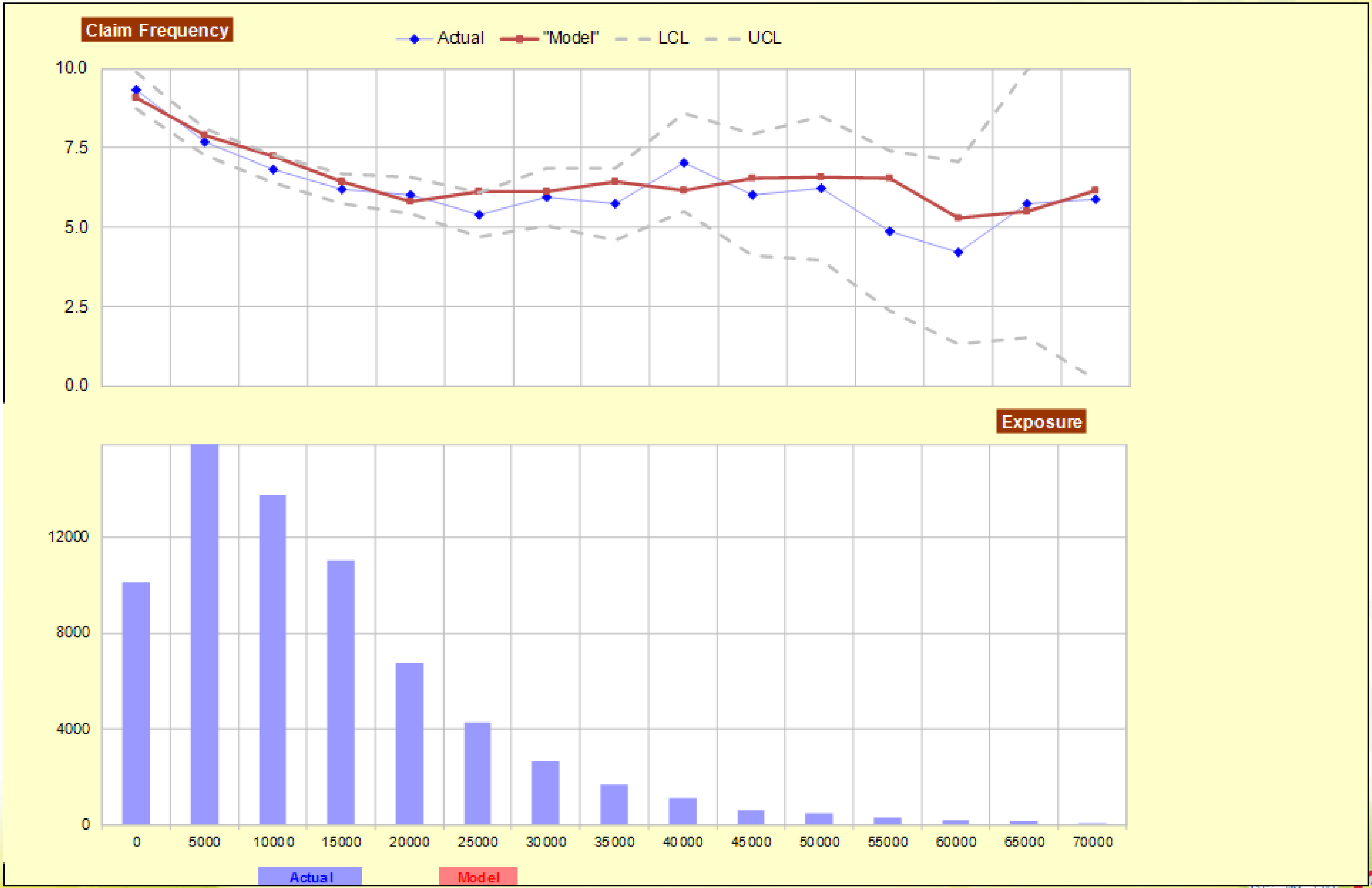
$$g(\mu) = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p + e_i$$

  - Y --- n x 1 Vector (measured), belonging to Exponential Family (Poisson, Gamma, Normal, Binomial, Inverse Gaussian, Negative Binomial, Tweedie)
  - Var(Yi) = f (E[Yi])
  - η --- η = Xβ
  - β --- p x 1 Vector (to be estimated)
  - X --- n x p (Design Matrix)
  - g --- Link Function

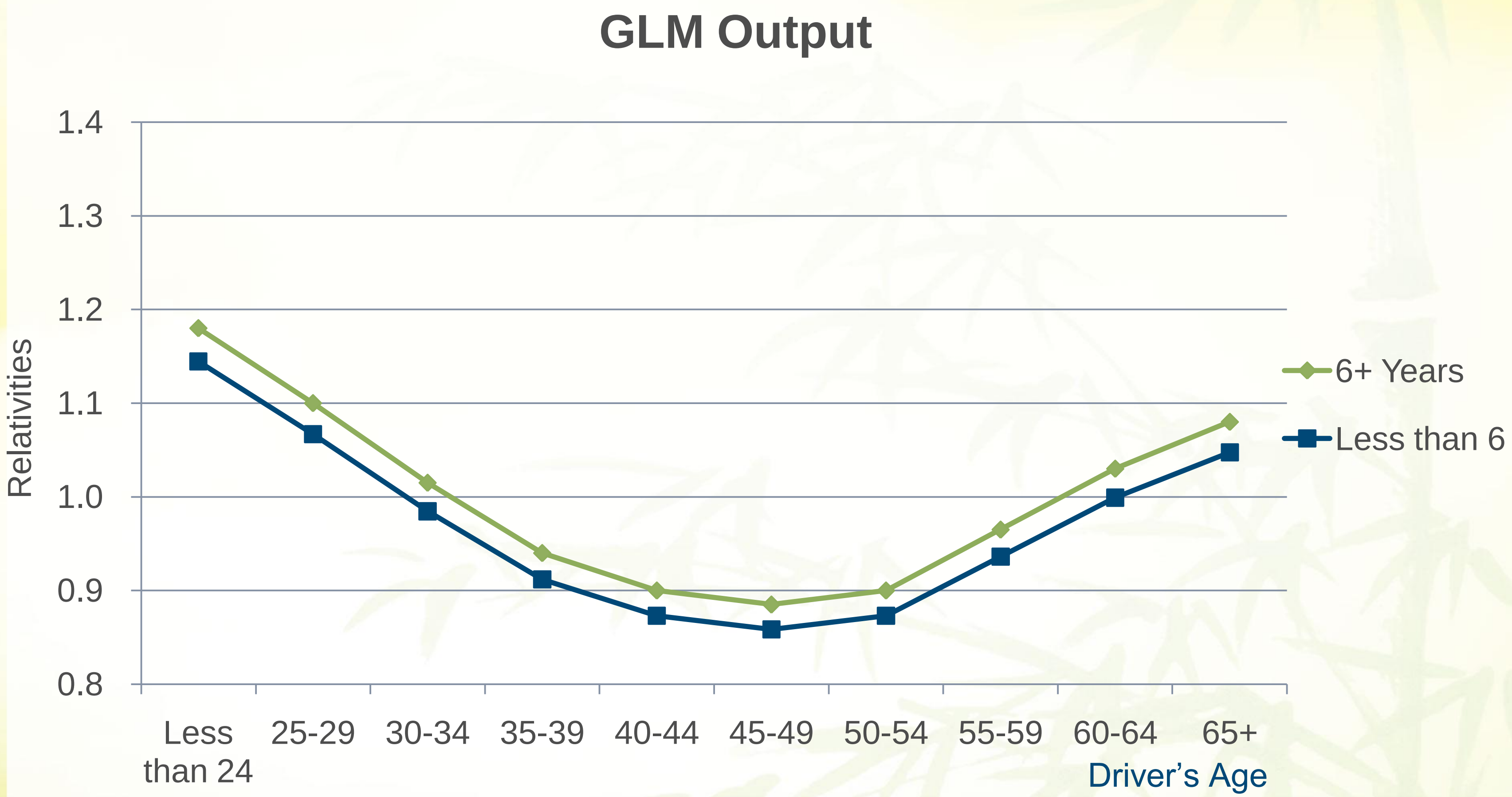
GLM: What is a Good Model?

- Consistent over time and withstand random sampling tests
- Strikes a balance between fitting well and over-fitting the data
- Various measures and tests can be done using a combination of :
  - AIC/BIC
  - Residual plots
  - Cramer's V - test the correlation of two categorical factors
  - Deviance
  - Chi-square
  - Confidence interval of fitted values for each factor
  - Gini

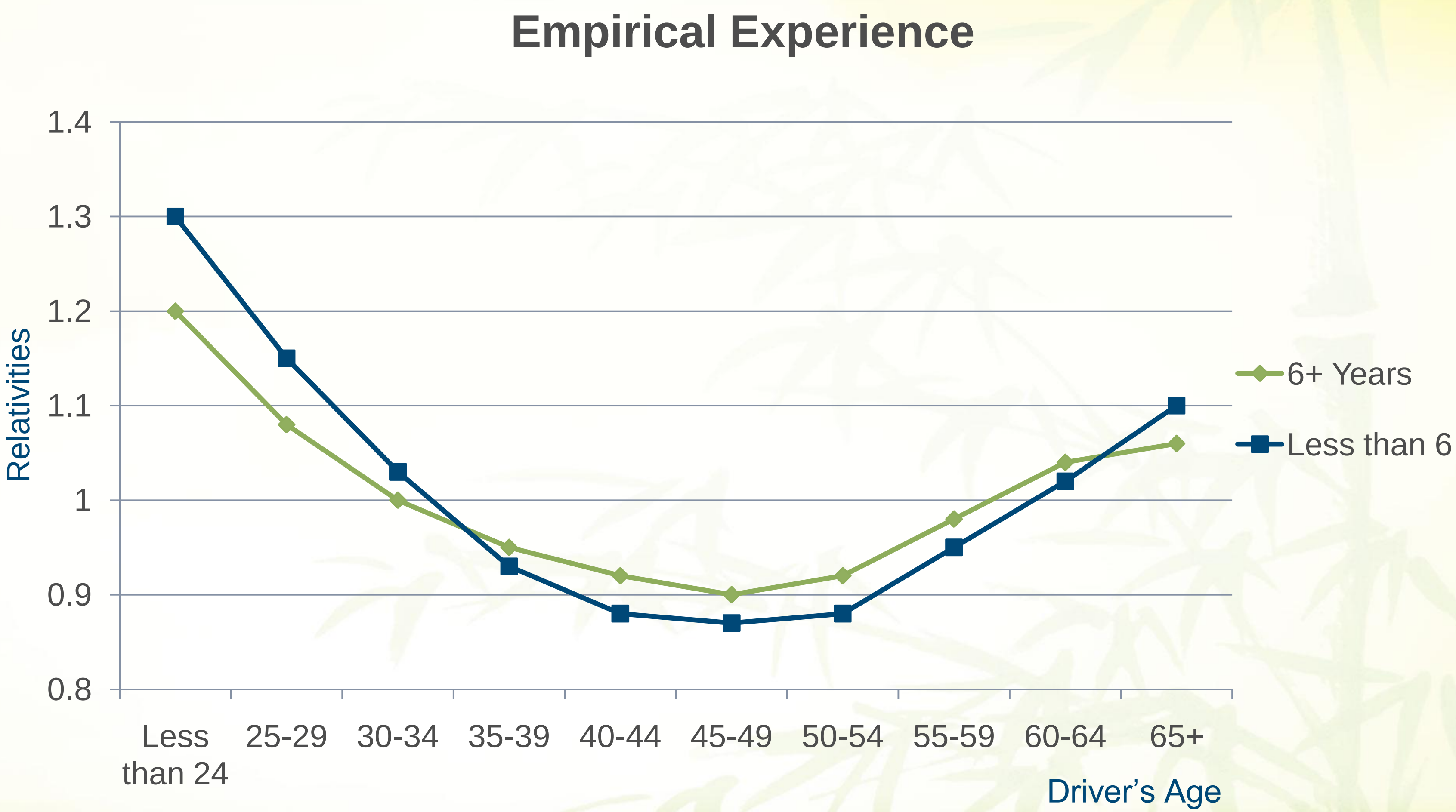
GLM: Revealing the Risk Shape



GLM Results: Does the Curve Fit?



GLM Results: Does the Curve Fit?



MACHINE LEARNING

Machine Learning

What is Machine Learning

- *"a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases." Wikipedia*

Addresses the fundamental questions directly

- Where are we making money?
- Where are we losing money?
- Can we be confident?

Identifies risk segments that are credible and produce consistent results from year to year

- Iterative, artificial intelligence process
- User defines the degree of credibility within segments

Results are only as good as the algorithm

- A good algorithm will maximise the number of segments identified

Technology/Modern Statistical Techniques is the Differentiator…

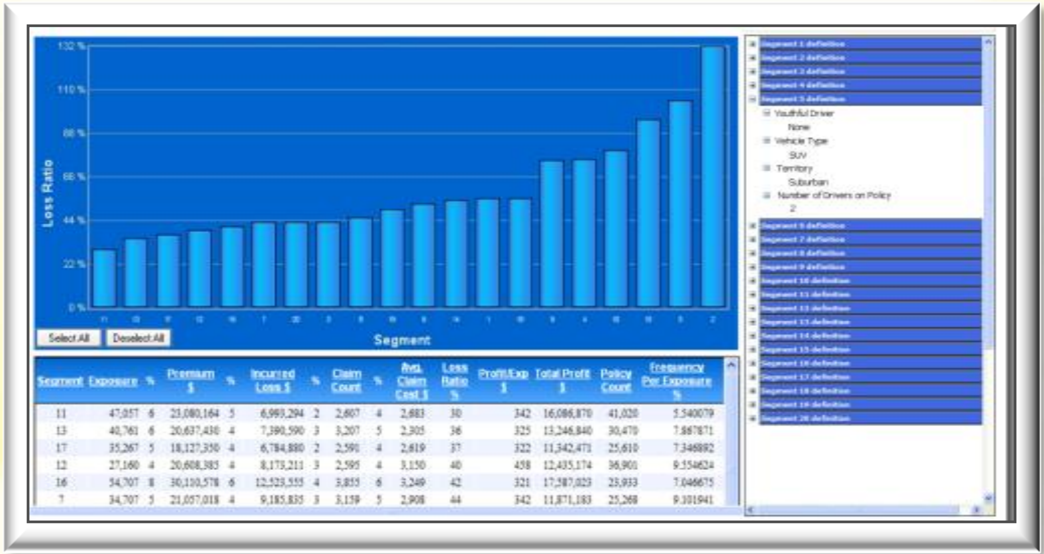
Current Methods (like GLM/GAM)

- ✓ Assumes that variables are independent *unless specifically defined otherwise*
  - ✓ "Optimal" predictors are based on assumptions
  - ✓ Can't solve what you don't know
- ✓ The number of risk attribute/value interactions is too large for a human to investigate given real-world resource and time constraints, therefore only a very small subset is investigated
- ✓ Pricing models are done at a coverage level versus a customer level

Machine Learning (like Ensembles)

- ✓ Allows data to interact naturally to find the patterns between characteristics within the data
- ✓ Finds the trade-off between over- and under-fitting automatically
- ✓ Does not require the user to specify the predictors and interactions to be included in the model - it discovers them!
- ✓ Extremely Fast and Efficient
- ✓ Performed at coverage, unit, or policy level

EEA Segmentation Analysis Types



Typical Uses:

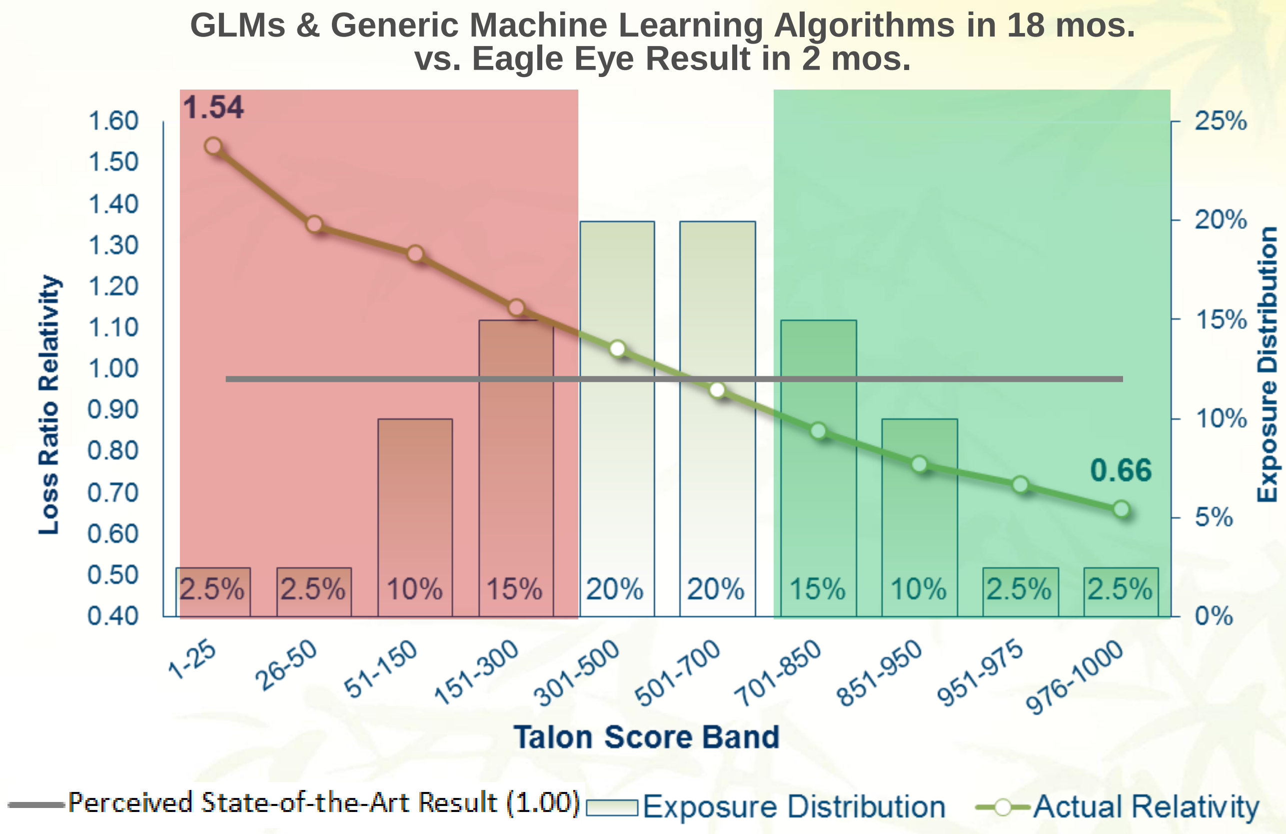
- Rate plan improvement
- Underwriting rules
- Target marketing

✓ Partitions the whole "universe" into exhaustive and mutually exclusive segments

✓ Available model responses: loss ratio, pure premium, frequency, severity, profit, retention

✓ Segments:
- Described by significant attributes
- Plain English description, easy to understand and actionable
- "Complex" compound variables

Pricing & UW: Find Errors



Talon vs. the perceived state-of-the-art solution:

We found 60% of the exposures in their technical premiums had pricing errors greater than 10%.

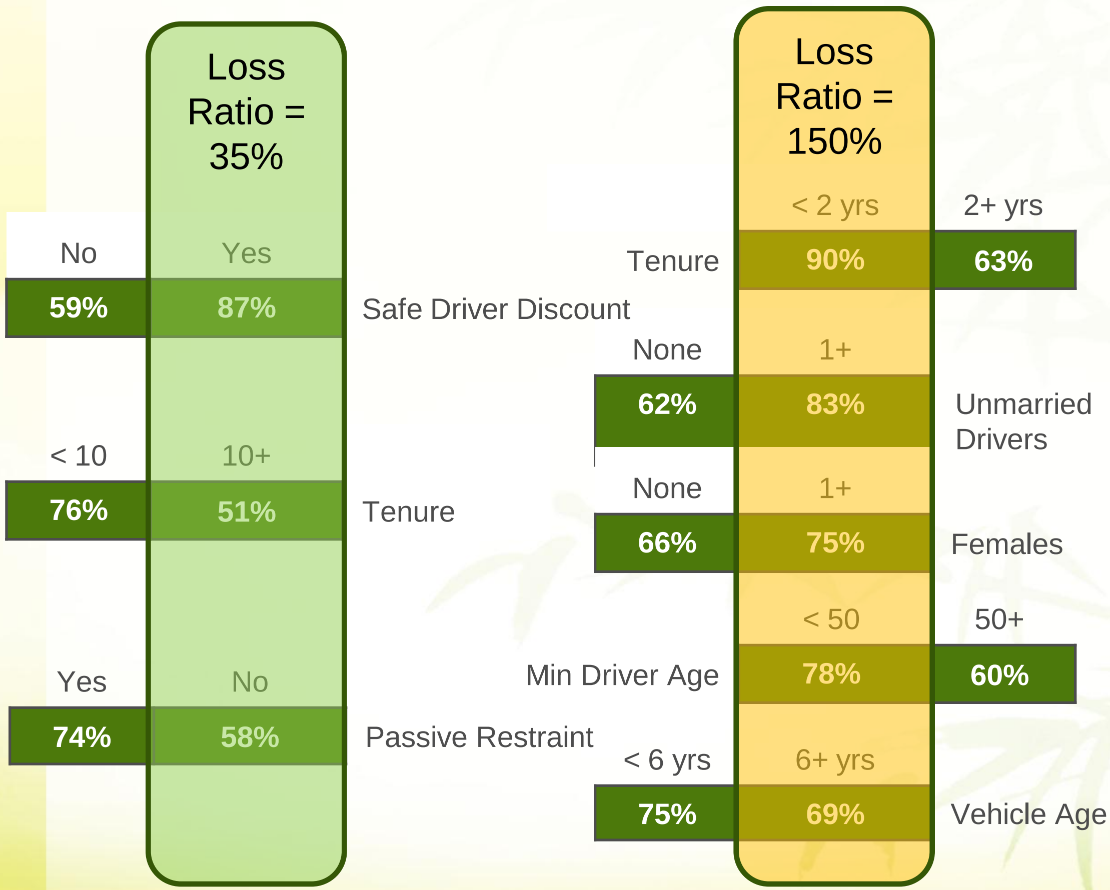
- Underpricing errors of up to 54%
- Overpricing errors of up to 34%

The difference: Talon’s learning algorithms are designed specifically for insurance data.

Price & UW – How Talon finds the Errors

Private Passenger Auto
(Total Portfolio Loss Ratio = 71%)

Best Customers *Worst Customers*



Identify New Patterns in the Data

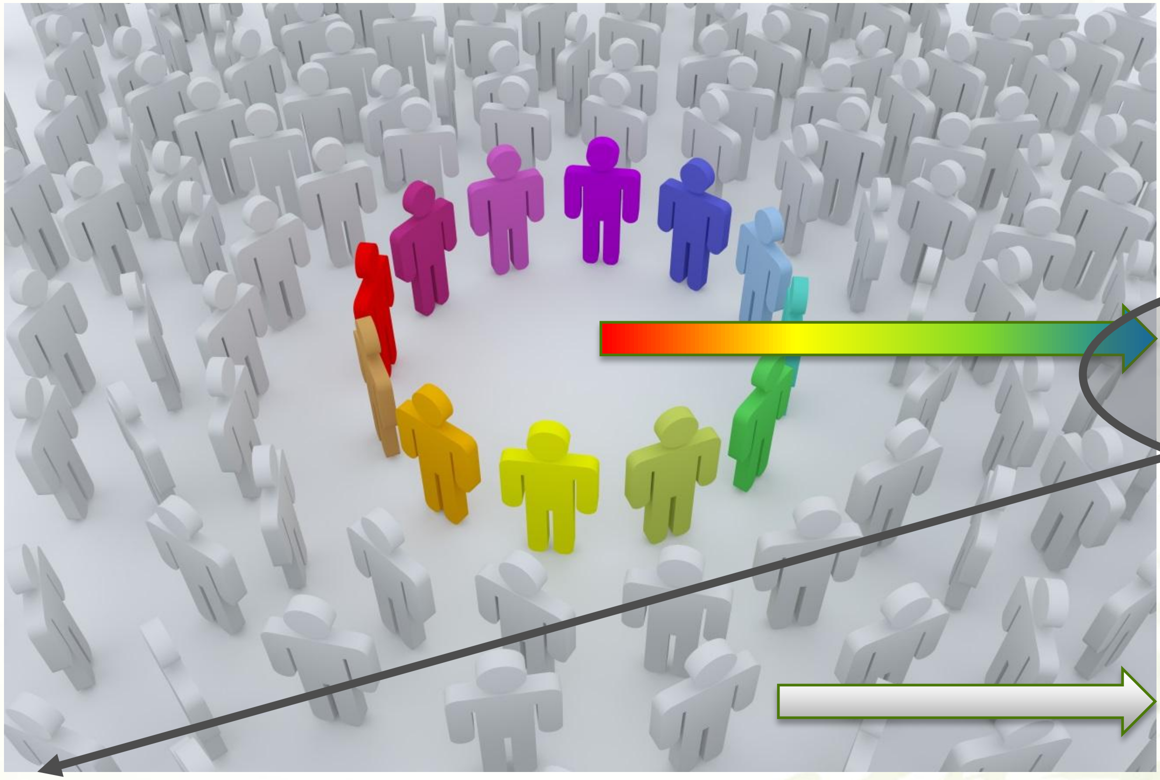
Some of the best customers are overpriced
Unique Pattern: Combining
- Safe Driver Discount
- 10 year old policy, or older
- No Passive Restraints

produces lowest loss ratio of 35%, 36 points lower than carrier average.

Some of the worst customers are Underpriced
Unique Pattern: Combining 5 unique variables, including customer tenure, marital status and vehicle age, identifies unprofitable business with loss ratios at 2x carrier average.

Pricing - Main Concept for Telematics

- Rapid Pricing Diagnostics using Machine Learning:



Combine regular policy pricing with Telematics data analysis*:

This needs new modeling technology!

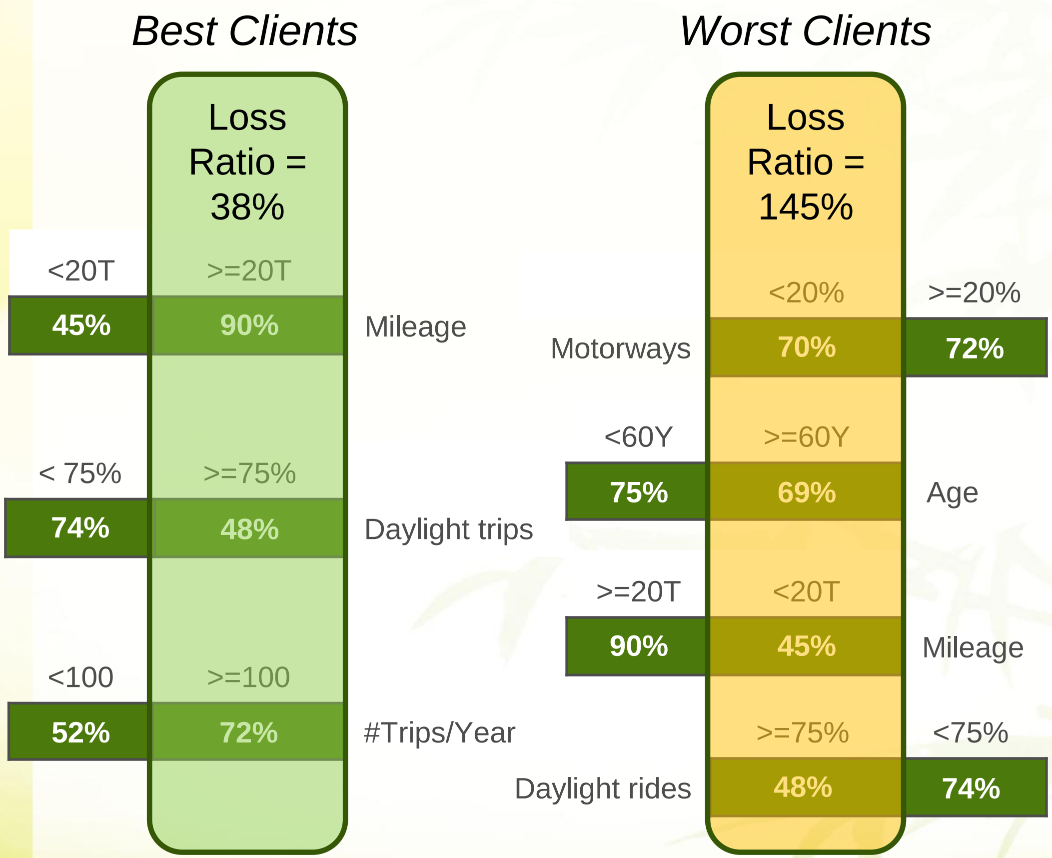
Price Difference = Loss Ratio = Telematics Claims / TP (Telematics)

Technical Price = TP (Standard Policy)

*Such analysis cannot be done with classical methods like GLMs because
a) Cost effecting, complex interactions within the Telematics data can only be detected automatically (through Machine Learning)
b) The price difference cannot be fitted by a GLM-Distribution

Pricing – Machine Learning for Telematics

Auto Telematics Product
(Total Portfolio Loss Ratio = 106%)



Identify very profitable and unprofitable segments

Some of the best customers might be overpriced
Unique Pattern: Combining

- High Mileage
- Mostly Day light
- Many trips

produces lowest loss ratio of 38%,

Some of the worst customers are underpriced and might be unexpected from their univariate patterns and can lie close to their good counterparts!

CASE STUDY

FROM MODELS TO RESULTS

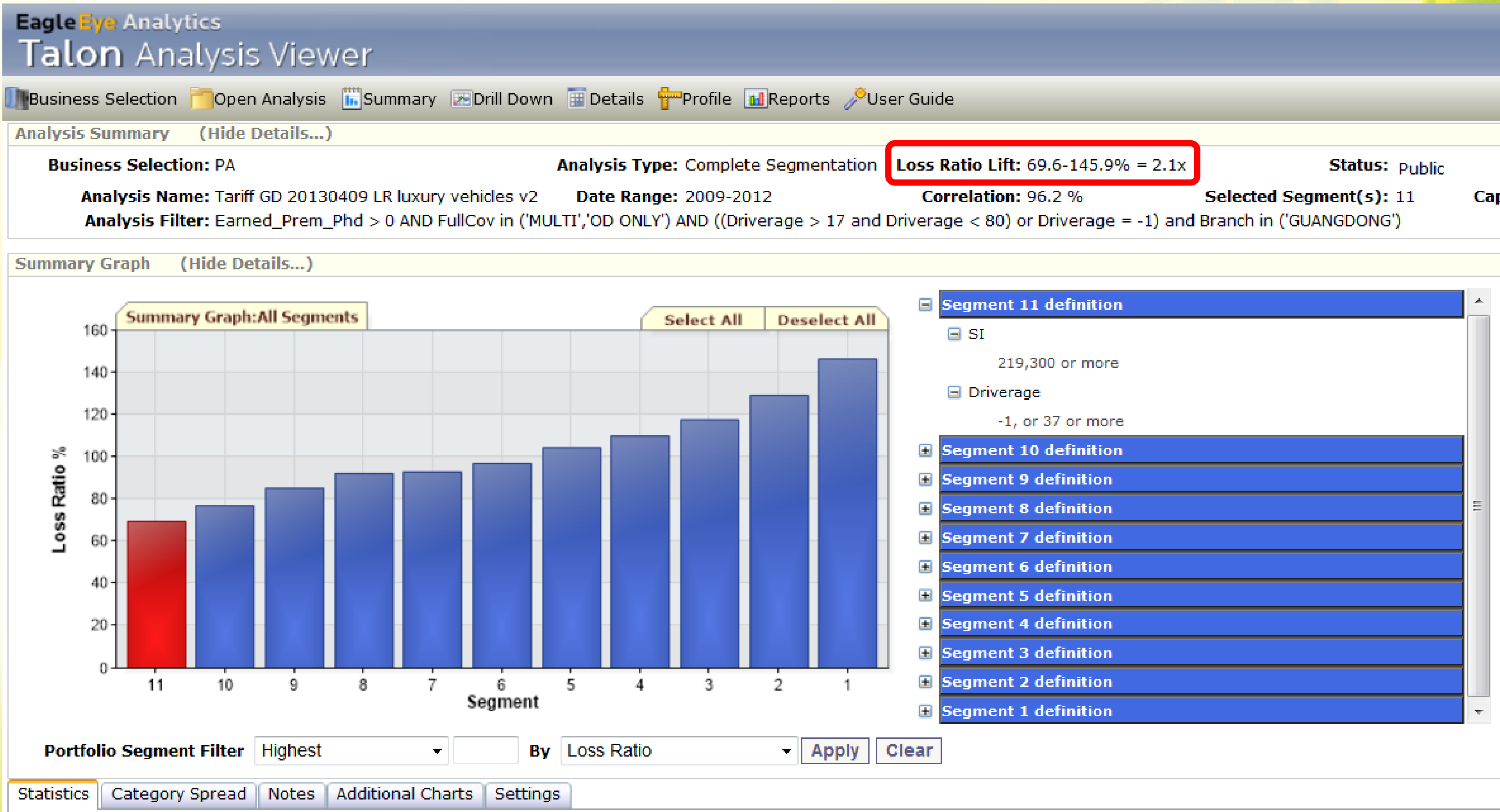
China Motor Tariff

Private Vehicle	Own Damage							
	Less than 1 year		1-2 years		2-6 years		6+ years	
	Fixed premium	Rate (%)	Fixed premium	Rate (%)	Fixed premium	Rate (%)	Fixed premium	Rate (%)
<6 Seats	539	1.28	513	1.22	508	1.21	523	1.24
6-10 Seats	646	1.28	616	1.22	609	1.21	628	1.24
10+ Seats	646	1.28	616	1.22	609	1.21	628	1.24

Rating Factor		Factor	Description
Designated Driver C1	C1a	0.9	With designated driver
	C1b	1	Undesignated driver
Driver's Age C2	C2a	1.05	Younger than 25 years old
	C2b	1	[25,30)
	C2c	0.95	[30,40)
	C2d	1	[40,60)
	C2e	1.05	At least 60 years old
Gender C3	C3a	1	Male
	C3b	0.95	Female
Driving Years C4	C4a	1.05	Less than 1 year
	C4b	1.02	[1,3)
	C4c	1	At least 3 years
Pre-defined Driving Area C5	C5a	1	Domestic
	C5b	0.95	Within province
	C5c	0.92	Routine
Average Annual Mileage C6	C6a	0.9	ILess than 30,000 km/p.a.
	C6b	1	[30000,50000) km/p.a.
	C6c	1.1-1.3	At least 50,000 km/p.a.
Traffic Violation Record C8	C8a	0.9	No liable traffic ticket record in previous year
	C8b	1	Liable traffic ticket record (s) in previous year

Significant Improvement on the Tariff

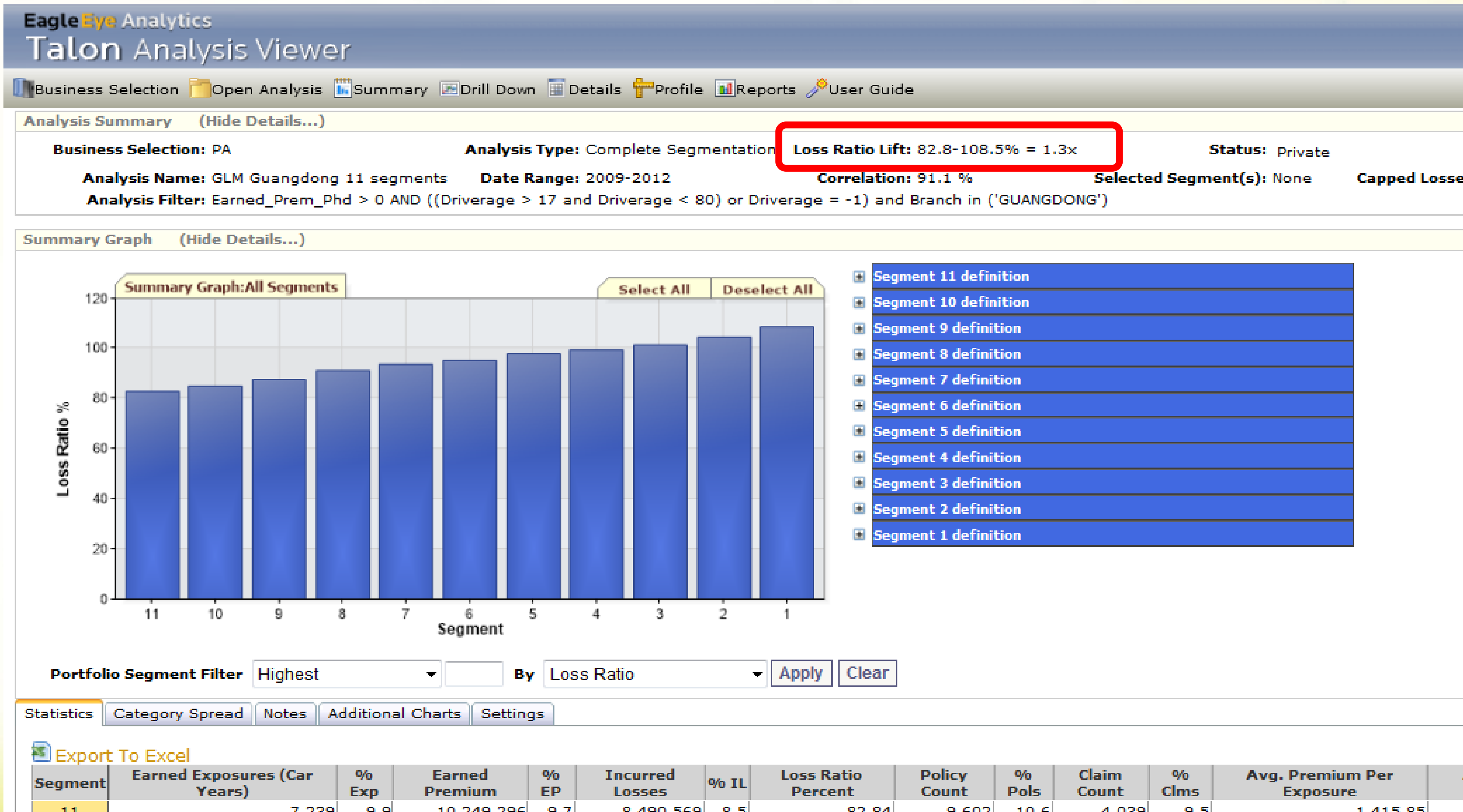
A Southern China Branch



Significant Improvement on GLM

A Southern China Branch

- High loss ratio means GLM underpriced and vice-versa
- Low loss ratio means GLM over-priced



Segmentation Result Drill Down Worse Segment

Talon Analysis Viewer

Business Selection | Open Analysis | Summary | Drill Down | Details | Profile | Reports | User Guide

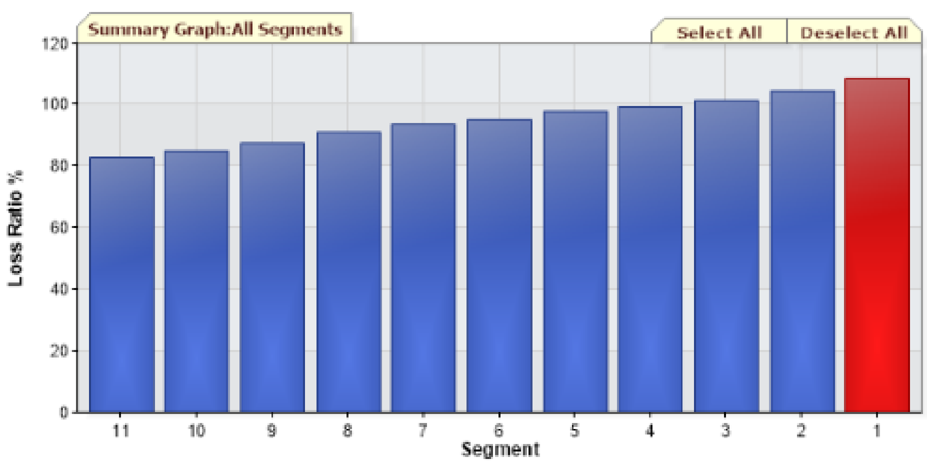
Analysis Summary (Hide Details...)

Business Selection: PA **Analysis Type:** Complete Segmentation **Loss Ratio Lift:** 82.8-108.5% = 1.3x

Analysis Name: GLM Guangdong 11 segments **Date Range:** 2009-2012 **Correlation:** 91.1 % **Selected Se**

Analysis Filter: Earned_Prem_Phd > 0 AND ((Driverage > 17 and Driverage < 80) or Driverage = -1) and Branch in ('GUANGDONG')

Summary Graph (Hide Details...)



Segment 2 definition
Segment 1 definition
Branch2
0201
0202
0203
0205
0206
0207
0208
0209
0210
0211
0212
0213
0214
0215
0216
0217
0218
0219
0220

Portfolio Segment Filter Highest **By** Loss Ratio Apply Clear

Statistics | Category Spread | Notes | Additional Charts | Settings

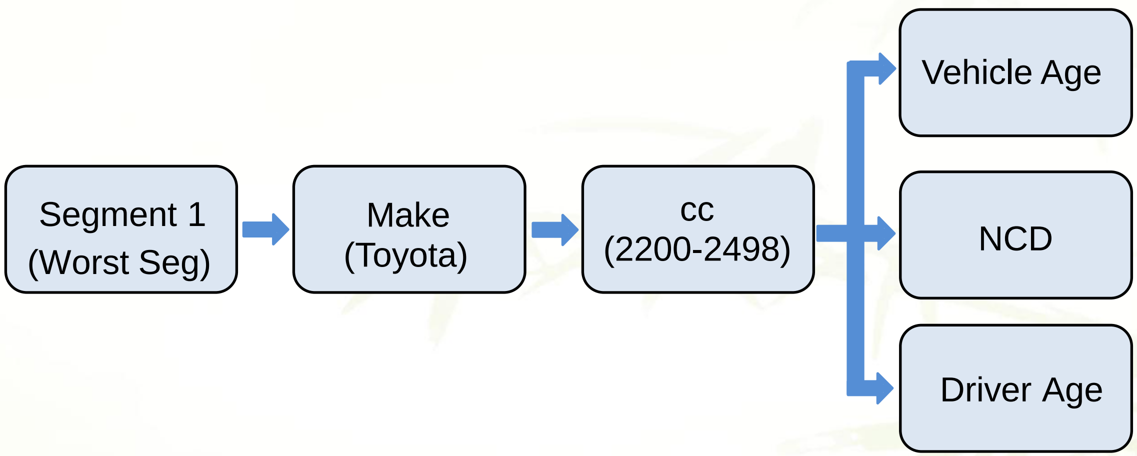
Export To Excel

Segment	Earned Exposures (Car Years)	% Exp	Earned Premium	% EP	Incurred Losses	% IL	Loss Ratio Percent	Policy Count	% Pols	Claim Count	% Clms
11	7,239	9.9	10,249,296	9.7	8,490,569	8.5	82.84	9,602	10.6	4,039	9.5
10	8,238	11.3	13,351,755	12.7	11,362,421	11.4	85.10	10,029	11.1	4,471	10.5
9	6,319	8.7	8,230,383	7.8	7,198,459	7.2	87.46	7,958	8.8	3,527	8.3

Segmentation Result Drill Down
Worse Segment

- Worse Segment here means GLM has Under Priced the risk

Segment	Branch2	Driver Age	SI
1	Over 20 branches in this segment!	18 to 35 (inclusive)	> RMB 100,000



All Ages Poor

Most exposure in 0%-10% NCD

Most ages but particularly 25-29

SUMMARY

Improvement on GLM – Using ML Results

- Introduced new interactions based
- Introduced new rating variable

Results

✓AIC Improved

✓BIC Improved

✓Gini Improved

✓Chi-square – just as good

✓Deviance Improved

Recap

- Need for Predictive Modelling is Today
- GLM is a robust pricing approach

BUT

- Machine Learning will
  - Tackles GLM's shortcomings
  - Identify critical hidden "gems" and "pitfalls"
  - Speed up the model build process systematically

The Value

Most Predictive Signal

- Lift curves of 2x-4x or more over other methods
- 4-6 way or more data interactions
- Non-linear interactions
- Local effects
- High correlations, over 90%

Fast

- Hundreds of iterations produced in a few hours
- Results in 60-90 days
- Real-Time Scoring Service supports real time decision-making

Actionable

- Understandable Segments & Scores
- Forward looking Management tools for Enterprise-wide application
- Approved rate filings in regulated markets

Why EagleEye Analytics?
From the perspective of clients

Most powerful and actionable predictive signal

After a failed attempt at getting a different, larger multivariate software provider to produce results specific to our company, we abandoned them. With Talon we have already Implemented the model results and are seeing the changes come to fruition.”

Speed to business impact allowing for real time excellence

Talon is extremely fast and efficient. It allows us to process analyses in a matter of minutes or hours. We now have the ability to quickly implement and maintain a sustainable competitive advantage.”

Complete vision

“EagleEye Analytics” solution suite gives us a common platform from which to dialogue regarding analytics and business performance throughout the enterprise into such areas as pricing, underwriting, claims and marketing. We now have a common and robust analytical foundation being used across our entire portfolio by multiple constituencies.

Proven results

We correctly determined that the cost of not utilizing EagleEye’s solution suite was too great to ignore. It is the most innovative, unique and powerful approach to driving profits, avoiding adverse selection, and improving our competitive advantage.”

2014 EAAC IN TAIPEI TAIWAN

18TH EAST ASIAN ACTUARIAL CONFERENCE

12-15 October 2014

Taipei International Convention Center in Taipei Taiwan

Questions？

EAAC