Using Novel Data for Vehicle Rating

Lakshmi Shalini and Mark Richards

CAS Special Interest Seminar: Baltimore, October 2011

SM

MEASURE, MANAGE, & REDUCE **RISK**

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.

Outline

1. Vehicle Characteristics vs. Series
2. Collecting and attaching data
3. Developing and Implementing Models
4. Some illustrative results

Vehicle Series

Working Definition: A vehicle series is an collection of vehicles that shares a number of characteristics in common and is used to aggregate loss experience.

- Different companies or organizations will partition the universe of vehicles in different ways, so the specific set of series will be similar across organizations but not identical.

Vehicle Series

- Common aggregations include:
  - Model year
  - Make
  - Model name
  - Additional attributes include:
    - Body Style &/or # of doors
    - # of drive wheels
    - Engine
    - Trim packages.

- Multiple price points (MSRPs) within series sharing common experience may lead to further refinement.

Vehicle Series

sounds simple but…:

- Model year (or range of model years).
  - ***When does the design change "significantly" enough to warrant a new series?***
- Make (manufacturer).
  - ***Chevy vs. GMC (~~Oldsmobile~~, ~~Pontiac~~, Buick, Cadillac)?***
- Model name (or aggregations like truck weight class).
  - ***VW Jetta / GTI / Fox / Golf?***
  - ***Ford Escape vs. Mazda Tribute?***
- Additional attributes, …
  - ***Irrelevant alternatives?***

…Credibility? …

Vehicle Characteristics

Alternate approach:

• Instead of defining a series, *link the loss experience* ***directly*** *to the characteristics of the vehicle*.

• Let a model *discover* the relationship between claims and the *relevant* aspects of a vehicle:

Model year	Price	Body style
# of doors	# of cylinders	# of drive wheels
Displacement	Horsepower	Torque
ESC	ALB	DRL
Curb weight	Wheelbase	***etc.***

Vehicle Characteristics

- *When does the design change "significantly" enough to warrant a new series?*
  - ***When / as much as the characteristics do.***
- *Chevy vs. GMC (~~Oldsmobile~~, ~~Pontiac~~, Buick, Cadillac)?*
  - ***The relevant differences are the characteristics, not the nameplate.***
- *VW Jetta / GTI / Fox / Golf?*
  - ***Design changes are considered, "branding" isn't.***
- *Ford Escape vs. Mazda Tribute?*
  - ***Share platform and common attributes, but some differences exist and are accounted for.***
- *Irrelevant alternatives?*
  - ***Not significant in models.***

Proxies vs. Characteristics

Proxies (working definition): attributes that are correlated with other relevant factors.

• Some of the relevant factors may be known, some may be readily available and others may not be easily measured or obtained.

• Proxies in models or series ratings may reflect or approximate the relationships inherent in the correlated factors, *but do so imperfectly*.

Proxies vs. Characteristics

Example: sedan with the same year, make and model.

Trim Level	Price (MSRP)	Horsepower	Braking Dist.
Base	$14K	120	X
Performance	$35K	276	0.8X

- Price captures the relationship between two performance measures that move in different directions.

Example: truck series from same make and year.

Truck Series	Price (MSRP)	Horsepower / Torque	Gross Weight
“15” (1/2 ton)	$21K	215 / 235	6,000
“25” (3/4 ton)	$28K	380 / 400	8,650
“35” (1 ton)	$36K	350 / 650	11,500

- Trucks are priced “by the pound” but also note that torque follows cost more closely than horsepower does.

Proxies vs. Characteristics

- Obtaining more detailed information (characteristics) can refine loss estimates that are approximated by proxies.
  - ✓ The proxy is still predictive in most cases
  - ✓ But, the magnitude of the effect is often dampened

- Other notable proxies:
  - ✓ Model year contains trends in engineering innovations
  - ✓ Model year is also correlated with price and miles driven

Collecting Data

In order to develop a model on vehicle characteristics, … **what data do we need?**

- Exposures and Losses at the specific exposure level.
- Other relevant rating factors (covariates):
  - Other applicable elements of the rating plan (Territory, Driver, etc.)
- Some vehicle specific characteristics (e.g. price, year, body style, # of cylinders, # of doors, etc.)

What data do we want?

- As much detailed, *relevant* vehicle specific characteristic data as we can *reasonably* get our hands on.

Where does detailed vehicle data come from?

- *A lot of hard work!*
  - …and multiple public and proprietary sources.

Obtaining 3rd Party Data

Outline

1. Qualifying data sources
2. Match keys
3. String matching tools
4. Level of aggregation
5. Process and QC

* Thanks to Leila Mortazavi of ISO Innovative Analytics and the team.

Qualifying Data Sources

- Is the data (*potentially*) predictive of losses?
- Is the data accurate? Can it be accurately matched?
- Completeness: does the data cover:
    - Adequate history (older model years)?
    - Adequately large proportion of insured vehicles?
- Will the data continue to be available in the future?
- Is the data allowable for use?
- Do you have (or can you obtain) appropriate rights of use?
- Does the data contain enough novel information to justify its cost (both the price and the time and effort to use it)?

Match Keys

Some working definitions:

- "***Base***" dataset: containing exposures, losses, covariates and vehicle VIN for the specific risk.
  - The match keys should be *at least* as refined (disaggregated) as the 3rd party data.
- "***3rd Party***" dataset(s): Multiple sources.
  - Different match keys and levels of aggregation.

- ***Ideally*** (i.e. unrealistically) we would be able to match all of our 3rd party data to our base data by VIN or some common *decoded* VIN.
  - *What follows is a discussion of what to do when the ideal situation doesn't hold.*

Match Key Cascade

Conceptually, the process of matching 3rd party data to the base can be thought of as hierarchical or a "cascade".

1. Model year
2. Manufacturer (Make)
3. Model Name
4. Body Style
5. Doors
6. Drive Wheels
7. Tie breakers *(data source specific)*

- ➢ If an exact match is found, then merge / join to base.
- ➢ If not, then roll up to next higher levels of hierarchy and resolve ambiguous cases.
- ➢ Hierarchy may differ for various 3rd party sources.
- ➢ Some pre-processing (clean-up) of keys helps a lot.

Match Key Details

1. **Model Year**: matches are relatively easy
   - Some sources provide data in model year ranges (e.g. 2003-2007).
2. **Manufacturer** (Make): also relatively easy
   - Differences easily resolved (e.g. 'ACUR' ⇔ 'ACURA')
3. **Model Name**: not easy at all – a great deal of source specific detail and some idiosyncrasies.
   - Some sources have two fields (e.g. "model" and "sub model").
   - Model names in one source can be parsed to create tie breakers (or keys) with a defined field in another source e.g.:
     - Drive wheels: "4X4" vs. "4X2", "AWD"
     - Engine type: "TURBO", "HYBRID", "FLEX"
     - Engine cylinders or displacement: "(V6)", "(V8)" or "2.0", "3.2"
   - Other differences / idiosyncrasies not easily resolved.
     - Some tools to aid in matching or disambiguation of model names will be described in detail below.

Match Key Details

4. **Body Style** …
5. **…**and **doors**: keep an eye out for differences

Base Data	3rd Party Data	
Body Style	**Body**	**Doors**
SEDAN 4D	SEDAN	4
COUPE 2D	COUPE	2
HCHBK 3D	HATCHBK	2

6. **Drive wheels**: ‘2’ or ‘ ‘ vs. ‘4’ (or ‘AWD’ or ‘6’)
7. **Tie Breakers**:
   - Common fields that exist across the base and 3rd party source (or that can be parsed from name).
   - Will differ from source to source.
   - Sometimes measurements differ slightly among sources (rounding, definitions) – need to accommodate differences.

String Matching Tools (in SAS)

SAS functions and routines

see: SAS 9.2 Language Reference: Dictionary, 4th Ed.

- SPEDIS: Spelling Distance [asymmetric]
  - Syntax: SPEDIS(query, keyword)
  - Performs a series of operations to convert “keyword” ➔ “query”
    - Assigns a cost to each operation, e.g.

Operation	Cost	Description
truncate	50	Delete a letter from the end
append	35	Add a letter to the end

  - Sums costs and divides by length(query) – rounds to nearest integer.
  - SPEDIS(string 1, string 2) not always equal to SPEDIS(string 2, string 1).

String Matching Tools (in SAS)

- COMPGED: Generalized Edit Distance
  - Similar to SPEDIS
    - Different operations & costs
    - More options
    - Doesn’t adjust for length
  - CALL COMPCOST: Use to modify (or ignore) operation costs in COMPGED
- COMPLEV: Levenshtein Edit Distance
- COMPARE: Position of leftmost character by which two strings differ
- SOUNDEX: Sounds Like
  - SOUNDEX(Couger) = SOUNDEX(Cougar)
- Also see: FIND, INDEX, etc.

Other software exists for evaluating string matches (e.g. Python).

String Matching Example

Base Model Name: "CAYENNESAWD"

3rd Party Model Names: "CAYENNETURBO"

"CAYENNE"

"CAYENNES"

Is the "best" match as obvious to the algorithm?

SPEDIS (CAYENNESAWD, CAYENNETURBO)

- Cost to convert CAYENNETURBO -> CAYENNESAWD
  - ***replace*** "TURB" with "SAWD" (cost to replace 4 = 100 x 4)
  - ***truncate*** "O" from the end (cost to truncate 1 = 50)
- **total cost = 40** = (400 + 50) / 11

SPEDIS (CAYENNESAWD, CAYENNE)

- Cost to convert CAYENNE -> CAYENNESAWD
  - ***append*** "SAWD" to end (cost to append 4 = 35 x 4)
- **total cost = 12** = 140 / 11

SPEDIS (CAYENNESAWD, CAYENNES)

- Cost to convert CAYENNES -> CAYENNESAWD
  - ***append*** "AWD" to end (cost to append 3 = 35 x 3)
- **total cost = 9** = 105 / 11

String Matching Example

Alternately, the Base Model Name: "CAYENNESAWD" could have been pre-processed to extract the "AWD" (into a tie breaker field):

- New Base MN: "CAYENNES", New Drive Wheels = "4" (or "A")

➢ then the SPEDIS example would be clear:

SPEDIS (CAYENNES, CAYENNETURBO)

- Cost to convert CAYENNETURBO -> CAYENNES
  - ***replace*** "T" with "S" (cost to replace 1 = 100)
  - ***truncate*** "URBO" from the end (cost to truncate 4 x 50 = 200)
- **total cost = 27** = (100 + 200) / 11

SPEDIS (CAYENNES, CAYENNE)

- Cost to convert CAYENNE -> CAYENNES
  - ***append*** "S" to end (cost to append 1 = 35)
- **total cost = 3** = 35 / 11

SPEDIS (CAYENNES, CAYENNES)

- Cost to convert CAYENNES -> CAYENNES
- **total cost = 0**

Matching Summary

- “Cascade” approach automates the discovery of exact matches and allows efforts to focus on disambiguation.
- A lot of pre-processing of fields is required to align them.
- String matching tools can aid in the process:
  - Each function has different aspects (costs, features and options).
  - Use multiple functions, and resolve disagreement (special cases).
- There is still a large manual effort.
  - EDA (Exploratory Data Analysis), data queries (group by, unique, …).
- Every different source requires unique solution details.
- The process needs to be replicable, in order to accommodate the introduction of new model years.

Aggregation in Data Sources

- Base data source should be as *disaggregate* as possible.
  - Merging / joining one row from a 3rd party source to multiple rows in the base is acceptable (and common).
  - Multiple rows in a 3rd party source matching a single row in the base is more *problematic*.
    - Are the differences in the rows of the 3rd party data source relevant (i.e. are they in fields that are not of interest / used in the model)?

Using 3rd Party Data

Process and Quality Control

- Initial matching process is very large:
    - > 25 model years.
    - > 100K distinct vehicles.
- Annual updates need to be executed quickly.
    - About 4,000 distinct vehicle make / model / trims per year.
    - Some percentage are new model introductions, some models are significantly redesigned , and some features are added / introduced or made standard equipment.
- A robust process with built in QC is required for the production process.

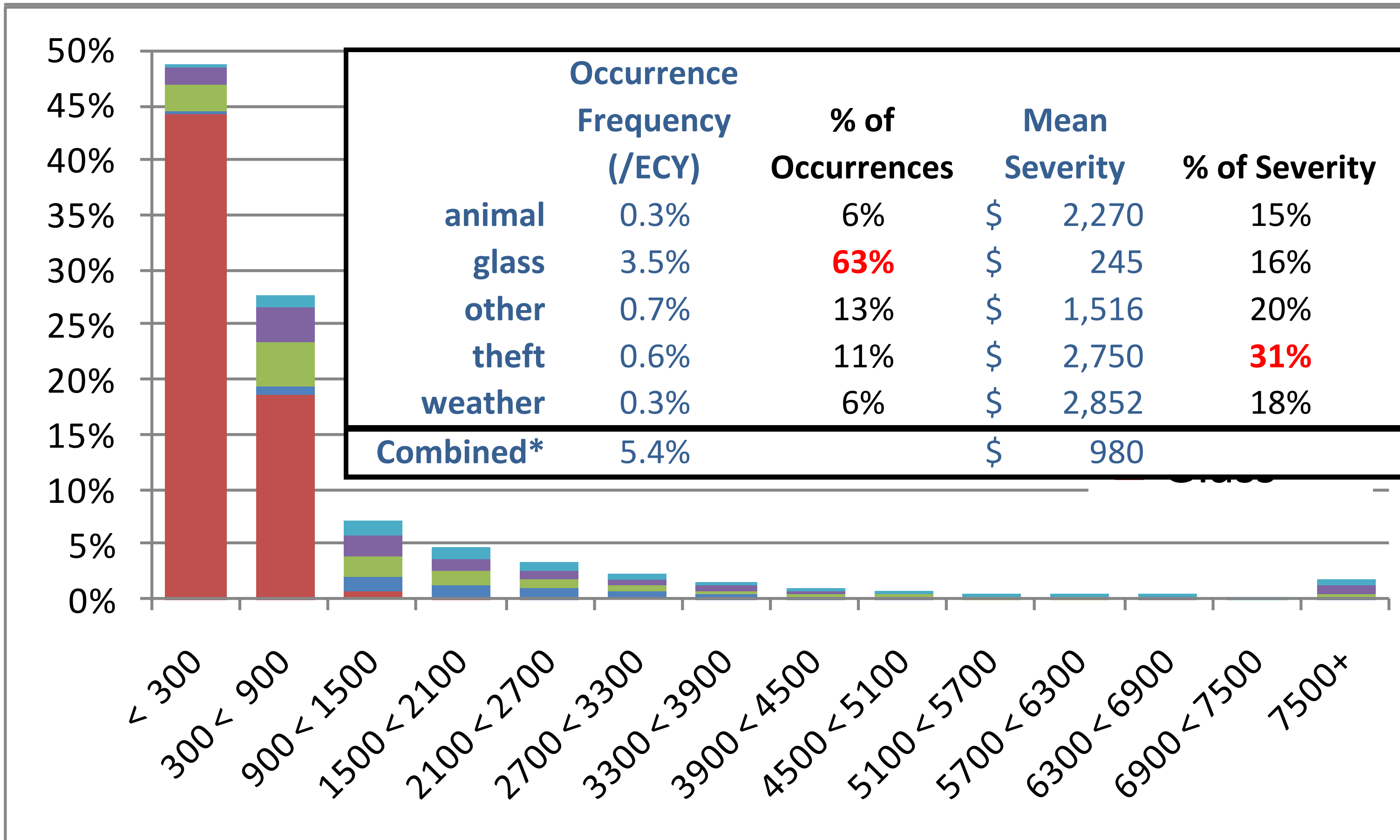
Developing Models

- When developing models from characteristics:
  - The variable selection task becomes challenging.
  - Need to adequately control for covariates (other elements of risk) like garaging address, driver, policy, etc.
  - Different characteristics may be associated with the likelihood (frequency) and the magnitude (severity) of losses, including antagonistic relationships (+/-).
  - Within a multi-peril coverage like comprehensive, different vehicle characteristics may be related to different perils.
    - The aspects of a vehicle that make it attractive to a thief may not matter to a deer.

Comprehensive Perils

By Peril - Frequency and Severity Distributions

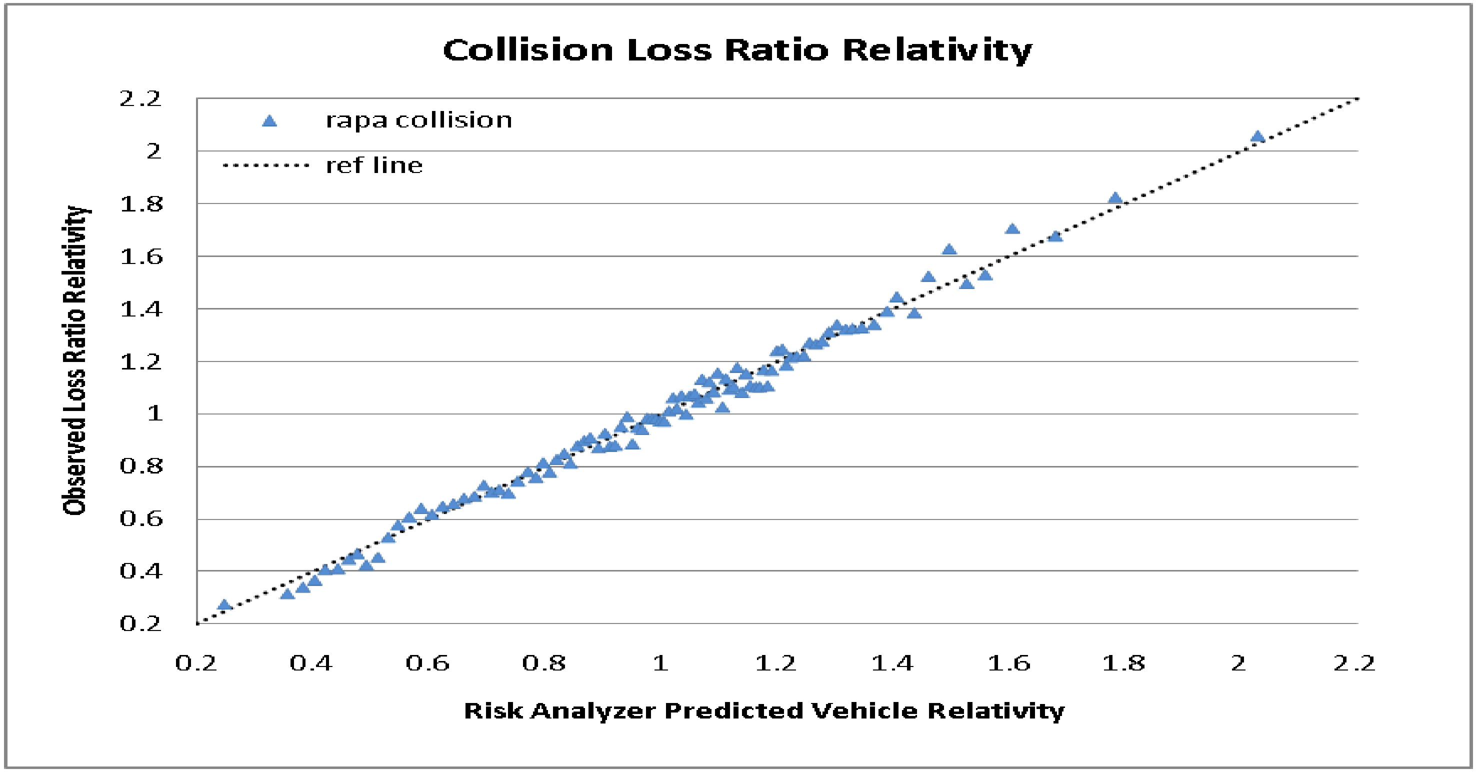
Comprehensive Losses – Severity Distribution



	Occurrence Frequency (/ECY)	% of Occurrences	Mean Severity	% of Severity	99th %tile Severity
animal	0.3%	6%	$ 2,270	15%	9,000
glass	3.5%	**63%**	$ 245	16%	967
other	0.7%	13%	$ 1,516	20%	9,388
theft	0.6%	11%	$ 2,750	**31%**	23,580
weather	0.3%	6%	$ 2,852	18%	18,901
Combined*	5.4%		$ 980		

Some illustrative results

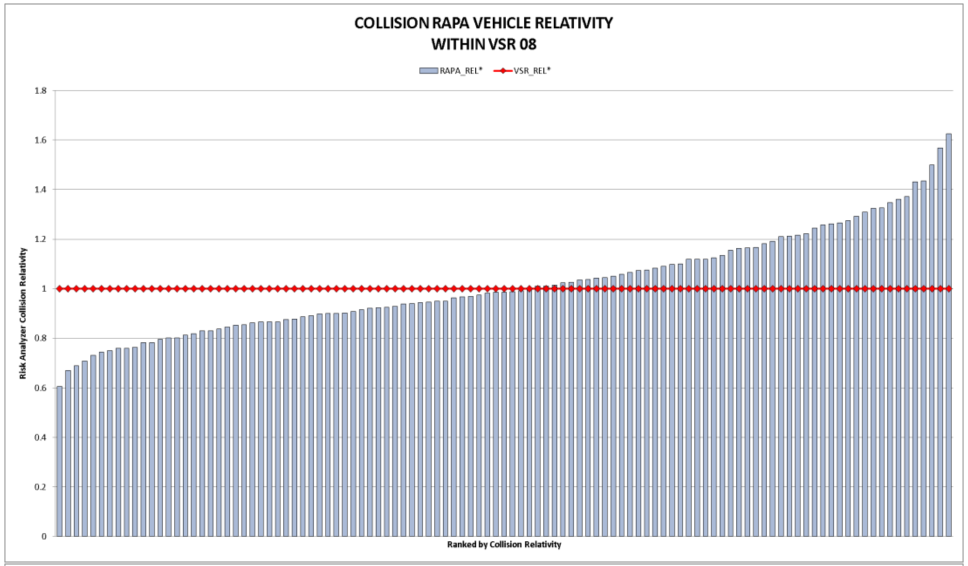
Collision Model Validation
Predicted Vs Actual



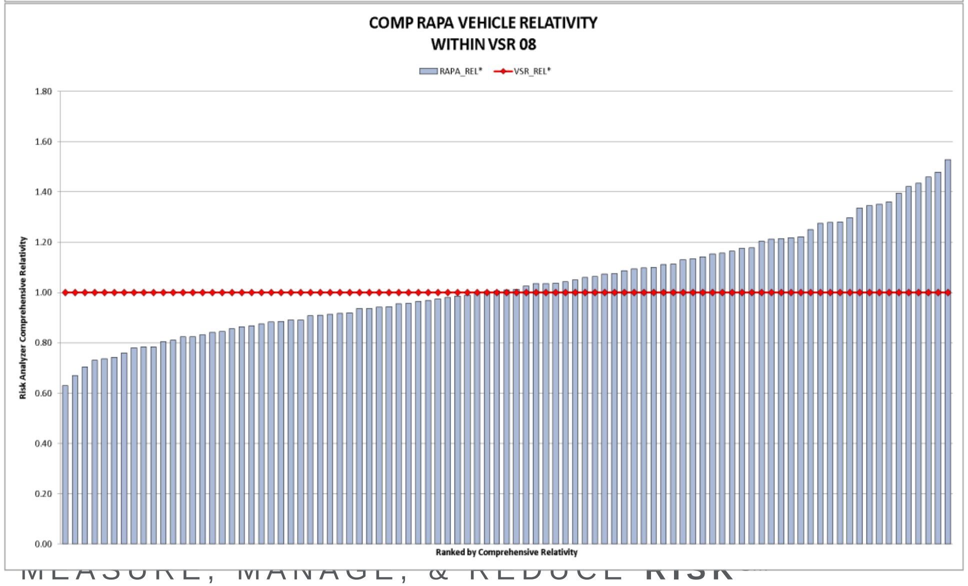
The vehicle model produces highly accurate predictions in line with the observed losses

Note: results against a holdout test dataset

Segmentation within VSR SYMBOL 08



Collision Coverage



Comprehensive Coverage

Predictive Modeling using Vehicle Characteristics provides significant segmentation within VSR Symbols

Example 1:Differentiation within series

2007 Ford Explorer Limited

Selected Attributes				RAPA Symbols/Relativities				VSR Symbols/Relativities		
Cyl	Horse-power	Curb-weight	Price New	COL SYM	COL REL	COM SYM	COM REL	SYM	COL REL	COM REL
6	210	4615	$34,070	LN	-	LJ	-	'12'	-	-
8	292	4615	$35,365	LP	+2%	LT	+19%	'12'	Same	Same

RAPA	➢RAPA Vehicle Module is able to pick up differences among several different styles of a common line, and differentiate the risks.
VSR	➢The VSR Symbol Set sometimes groups different model trims within a series together under a common VSR symbol.

Example 2: Performance Matters

2007 Honda Accord

Selected Attributes				RAPA Symbols/Relativities				VSR Symbols/Relativities		
Model Trim	Horse power	Engine Size	Cyl	COL SYM	COL REL	COM SYM	COM REL	SYM	COL REL	COM REL
EX	166	2.4L	4	HU	-	HT	-	'13'	-	-
SE	244	3.0L	6	HV	+5%	HV	+7%	'13'	Same	Same

COMP	➢The relativity for the EX model in RAPA is about 7% higher, compared to a 0% differential in VSR.

COLL	➢The relativity increase for the EX model in RAPA is about 5%, compared to a 0% differential in VSR.

Example 3: Redesigned Vehicle Series

Toyota Camry 4-Door SE

Selected Attributes			RAPA Symbols/Relativities				VSR Symbols/Relativities		
Model Year	Accel Rate	Price New	COL SYM	COL REL	COM SYM	COM REL	SYM	COL REL	COM REL
2006	X	$19,925	FR	-	FM	-	'11'	-	-
2007	1.6X	$18,270	EW	+15%	ER	+8%	'10'	-5%	-9%

COMP	➢The 2007 redesign produces an 8% ***increase*** in relativity over the prior version in RAPA. ➢Contrast with a 9% ***decrease*** in relativity in VSR

COLL	➢The 2007 redesign produces an 15% ***increase*** in relativity over the prior version in RAPA. ➢Contrast with a 5% ***decrease*** in relativity in VSR

Summary

- Vehicle series rating and vehicle characteristic driven modeling
- Techniques and challenges: vehicle data for modeling and results
- Questions?