# Multicollinearity and Sparse Data in Key Driver Analysis: Challenges and Solutions

**Presentation at the Predictive Analytics World Conference**
**Marriott Hotel, San Francisco** **April 15–16, 2013**

ISO 20252 Certified

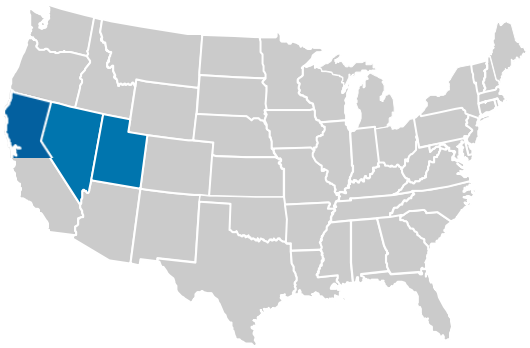
Ray Reno, Market Strategies International
Noe Tuason, AAA Northern California, Nevada, and Utah
Bob Rayner, Market Strategies International

# Agenda

- Background about AAA
- Analytical issues related to relative importance
- Example of two situations
  1. Shapley Value Analysis
  2. Bivariate regression for filtered events
- Conclusions and recommendations

# About AAA

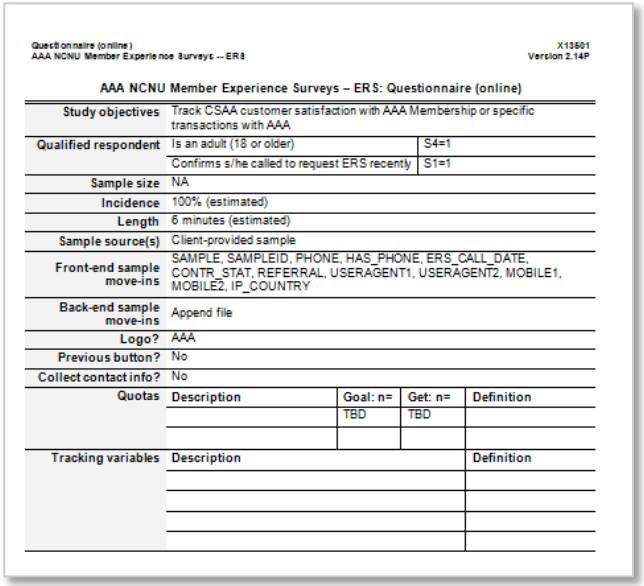
- We are a membership organization which provides different products and services to its members:
  - Emergency Roadside Service (ERS)
  - Insurance
  - Travel Services
  - Others
- Our company operates officially in Northern California, Nevada, and Utah (NCNU).
  But other AAA clubs are partners and affiliates and provide ERS to members regardless of club.



# Our promises to our members

# Are we keeping our promises?


- We conduct daily, weekly, and monthly surveys to get feedback from members on how we are doing with respect to our promises.



Questionnaire (online)
AAA NCNU Member Experience Surveys -- ERS

X13501
Version 2.14P

**AAA NCNU Member Experience Surveys – ERS: Questionnaire (online)**

| | | | | |
|---|---|---|---|---|
| Study objectives | Track CSAA customer satisfaction with AAA Membership or specific transactions with AAA | | | |
| Qualified respondent | Is an adult (18 or older) | | S4=1 | |
| | Confirms s/he called to request ERS recently | | S1=1 | |
| Sample size | NA | | | |
| Incidence | 100% (estimated) | | | |
| Length | 6 minutes (estimated) | | | |
| Sample source(s) | Client-provided sample | | | |
| Front-end sample move-ins | SAMPLE, SAMPLEID, PHONE, HAS_PHONE, ERS_CALL_DATE, CONTR_STAT, REFERRAL, USERAGENT1, USERAGENT2, MOBILE1, MOBILE2, IP_COUNTRY | | | |
| Back-end sample move-ins | Append file | | | |
| Logo? | AAA | | | |
| Previous button? | No | | | |
| Collect contact info? | No | | | |
| Quotas | Description | Goal: n= | Get: n= | Definition |
| | | TBD | TBD | |
| | | | | |
| Tracking variables | Description | | | Definition |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

- We look at their transactions with us: ERS, claims, purchase, renewals, upgrades, etc., to measure how we handle them.

- As a main measure of our service performance, we ask them how likely they would recommend us to their family and friends.  Responses to this question are the basis for the Net Promoter Score (NPS).

# Impact of key drivers in 2 key areas

**One of the goals of the analysis is to identify the potential drivers of NPS, and quantify their relative impact.**

## 1. Brand Image Drivers

- Company I can trust
- Meets their commitments to me
- Easy to do business with
- Good value for the money
- Keeps me safe and secure
- Knowledgeable Employees
- Products/services relevant to me
- Rewards me for my loyalty

**Recommend**

## 2. Transaction Satisfaction

- ERS Service
- Membership Renewal Process
- Insurance Renewal Process
- Membership Account Change Process
- Insurance Purchase Experience
- Insurance Policy Change Experience
- Membership Purchase Experience
- Insurance Claim Process
- Travel Purchase Experience

# Analytical issues

> On the face of it, with both the potential drivers and target using (11-point) metric scales, Multiple Linear Regression is the method to consider.

> However…

- The Brand Image Drivers are highly correlated.
- In the case where Transaction Satisfaction items are the key drivers of interest, most respondents experienced only one transaction. (In a sense, the data are composed of different respondents.)

# Analytical solutions

- Separate analytical approaches were used to deal with these separate issues.
- Shapley Value Imputation was used to calculate importance in the presence of multicollinearity.
- Regression coefficients were adjusted to estimate importance of different transactions when there were different samples responding to each.

# Shapley Value Imputation

## Assessing importance with regression and multicollinearity

- In practice, the relative importance of predictors (independent variables) in a regression model is frequently measured by the size of the standardized coefficients (or Betas) associated with each.
- The total influence of all the predictors in a model—*the amount of variance explained, or* $R^2$—is directly related to and can be viewed as a transformed measure of the summed predictor effects on the measure being explained.
  - If the predictors themselves are uncorrelated with each other, this is in fact the case.

# Shapley Value Imputation

## Assessing importance with regression and multicollinearity

- However, in market research— *as in all social science research which relies on inherently imprecise measures of complex concepts such as attitudes, orientations, judgments, etc.* — correlations among predictors, known as multicollinearity, is often the norm.
  - Multicollinearity when it is severe, results in imprecise and unstable coefficients and thus the relative importance among predictors cannot be accurately gauged.
  - Statisticians have developed a number of procedures to address the effects of multicollinearity.
  - Shapley Value Regression is one of the most recent methods (Lipovetsky & Conklin, 2001).

# Shapley Value Imputation

Shapley Value Regression has its origins in a Game Theory concept developed by Lloyd Shapley in the 1950s.

- Shapley was concerned with fair allocation of collectively gained profits between several collaborative actors. His main question was: *How can we "fairly" estimate the importance of each actor to the overall result given different amounts of contribution*?

- Shapley's solution to allocating importance fairly, which he formalized in a series of equations, has been borrowed and applied to the estimation of predictor importance in regression analysis when there is high multicollinearity.

# Shapley Value Imputation

## Shapley Value Regression Procedure

> This method puts Shapley's "fair allocation" of predictor importance to the final outcome into practice as follows:

1. Given a number of predictors, all possible combinations of those predictors are run against the final outcome (dependent variable). This includes each predictor by itself, each with each other predictor, each with pairs of others, and so on.
2. The average contribution to the $R^2$ of the model of each predictor (in all of its combinations) is computed. This averaged contribution becomes the importance measure for each of the predictors.

# Shapley Value Imputation

## Shapley Value Regression Procedure

> Calculation of average contribution to $R^2$

| Model | $R^2$ for predicting X |
|---|---|
| ABC | 0.5877 |
| AB | 0.5823 |
| AC | 0.4251 |
| BC | 0.5478 |
| A | 0.3306 |
| B | 0.5403 |
| C | 0.2442 |

| | X | A | B | C |
|---|---|---|---|---|
| X | 1.0000 | | | |
| A | 0.5750 | 1.0000 | | |
| B | 0.7350 | 0.5493 | 1.0000 | |
| C | 0.4942 | 0.3609 | 0.5757 | 1.0000 |

Shapley Value (A) = $[(R^2_{ABC} - R^2_{BC}) + [(R^2_{AC} - R^2_{C}) + (R^2_{AB} - R^2_{B})]/2 + R^2_{A})]/3$

$= [(0.5877-0.5478) + [(0.4251-0.2442) + (0.5823-0.5403)]/2 + 0.3306)]/3$

$= [0.0399 + (0.1809 + 0.0420)/2 + 0.3306]/3$

$= 0.1607$

# Shapley Value Imputation

## Shapley Value Regression Procedure

> Calculation of average contribution to $R^2$

| Model | $R^2$ for predicting X |
|---|---|
| ABC | 0.5877 |
| AB | 0.5823 |
| AC | 0.4251 |
| BC | 0.5478 |
| A | 0.3306 |
| B | 0.5403 |
| C | 0.2442 |

| | X | A | B | C |
|---|---|---|---|---|
| X | 1.0000 | | | |
| A | 0.5750 | 1.0000 | | |
| B | 0.7350 | 0.5493 | 1.0000 | |
| C | 0.4942 | 0.3609 | 0.5757 | 1.0000 |

Shapley Value (B) = $[(R^2_{ABC} - R^2_{AC}) + [(R^2_{AB} - R^2_{A}) + (R^2_{BC} - R^2_{C})]/2 + R^2_{B})]/3$

= $[(0.5877-0.4251) + [(0.5823-0.3306) + (0.5478-0.2442)]/2 + 0.5403)]/3$

= $[0.1626 + (0.2517 + 0.3036)/2 + 0.5403]/3$

= 0.3268

# Shapley Value Imputation

## Shapley Value Regression Procedure

> Calculation of average contribution to $R^2$

| Model | $R^2$ for predicting X |
|---|---|
| ABC | 0.5877 |
| AB | 0.5823 |
| AC | 0.4251 |
| BC | 0.5478 |
| A | 0.3306 |
| B | 0.5403 |
| C | 0.2442 |

| | X | A | B | C |
|---|---|---|---|---|
| X | 1.0000 | | | |
| A | 0.5750 | 1.0000 | | |
| B | 0.7350 | 0.5493 | 1.0000 | |
| C | 0.4942 | 0.3609 | 0.5757 | 1.0000 |

Shapley Value (C) = $[(R^2_{ABC} - R^2_{AB}) + [(R^2_{AC} - R^2_{A}) + (R^2_{BC} - R^2_{B})]/2 + R^2_{C})]/3$

$= [(0.5877-0.5823) + [(0.4251-0.3306) + (0.5478-0.5403)]/2 + 0.2442)]/3$

$= [0.0054 + (0.0945 + 0.0075)/2 + 0.2442]/3$

$= 0.1002$

# Shapley Value Imputation

## Shapley Value Regression Procedure

> *Continued*

3. Importance measures calculated this way are inherently reliably and stable.

4. Predictor importance computed this way results in relative importance values that sum to the $R^2$ from normal multiple regression with all predictors in the model.

Thus, it produces relative importances in a way that is a decomposition of $R^2$.

# Shapley Value Imputation

## Shapley Value Regression Procedure

> Calculation of average contribution to $R^2$

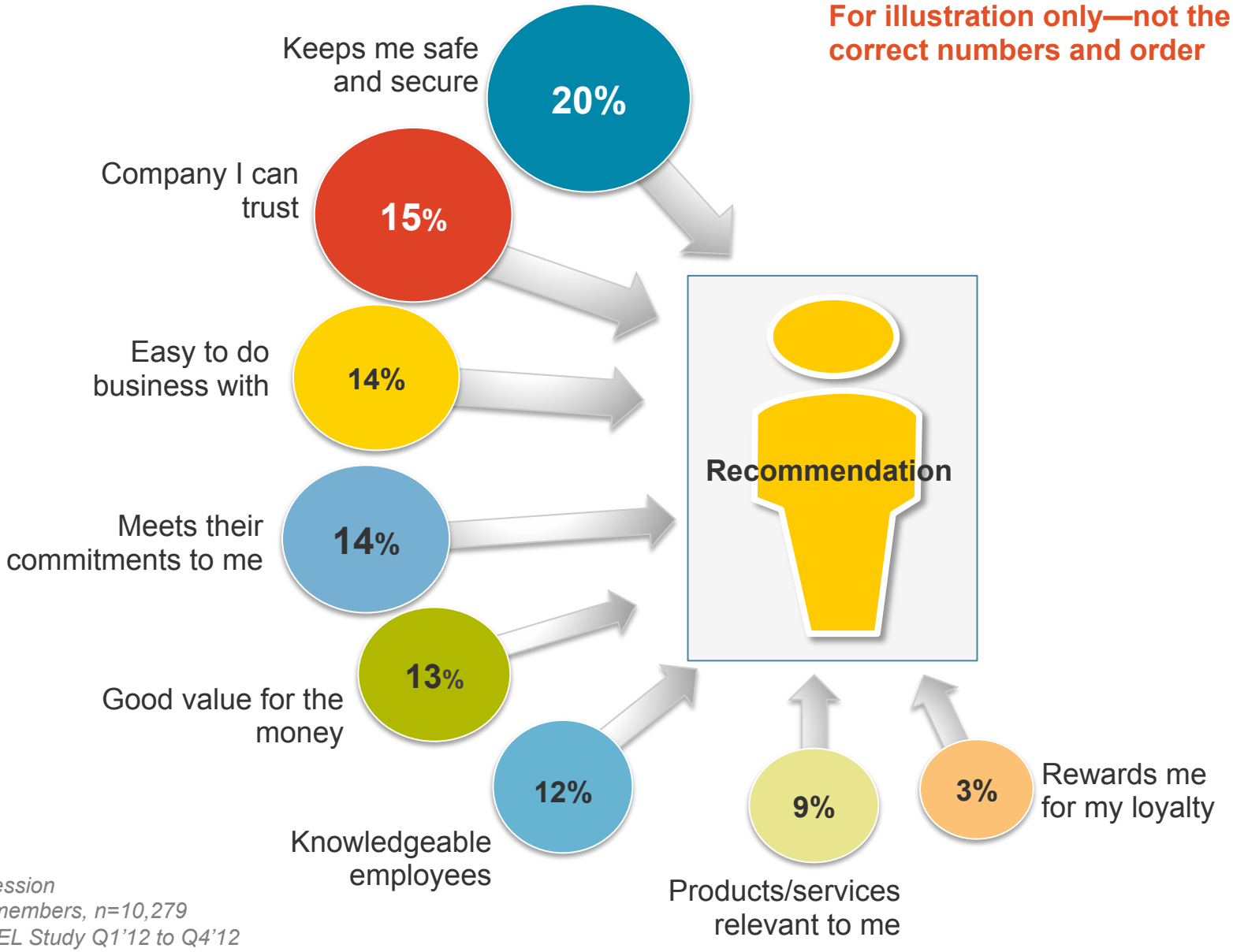
| Model | $R^2$ for predicting X |
|---|---|
| ABC | 0.5877 |

Shapley Value (A) = 0.1607
+
Shapley Value (B) = 0.3268
+
Shapley Value (C) = 0.1002

= 0.5877

# Relationship–Drivers of recommendation Image/ Brand Evaluations

**For illustration only—not the correct numbers and order**




**Shapley Regression*
*Base: All AAA members, n=10,279*
*Source: AAA REL Study Q1'12 to Q4'12*

# Adjusted bivariate regression

## Assessing importance with regression with filtering

- Conducting a traditional OLS Regression was problematic because respondents only provided a rating of satisfaction for the transactions they had.
  - This led to limited overlapping respondents across the different transactions.
    - Listwise regression approaches led to very small sample sizes.
    - Pairwise regression approaches led to unstable coefficients.

# Adjusted bivariate regression

## Assessing importance with regression with filtering

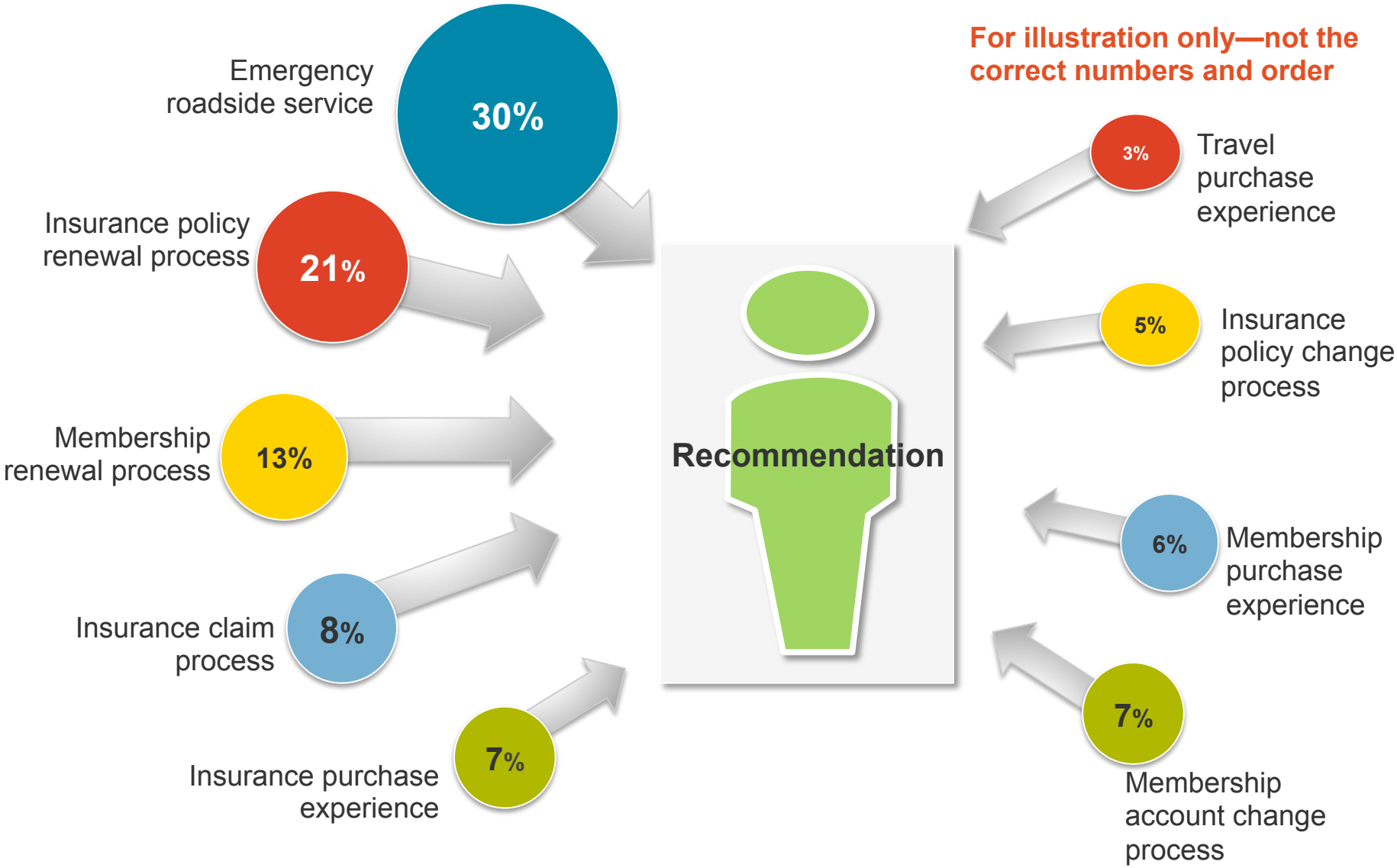
Followed a 2-step process of arriving at adjusted bivariate regression coefficients.

1. Run bivariate regressions of each predictor against *Likelihood to Recommend* to get an initial coefficient for each predictor.
2. Adjust for incidence of key driver.

| Event | Regression Coefficient | Incidence | Adj Reg Coefficient |
|---|---|---|---|
| Auto Claim | 0.47 | 11% | 0.05 |
| ERS | 0.47 | 44% | 0.20 |
| Home Claim | 0.32 | 2% | 0.01 |

# Relationship–Drivers of recommendation

## Transaction Satisfaction




*Arrow represents increase or decrease from previous year ranking in importance.
**Driver analysis conducted using binary regressions, regressing each predictor onto recommendation independently.
Relative importances reflect strength of each predictor's coefficient, relative to all other predictors.

# Conclusions and recommendations

- The two models presented here are only two of several models built by the company to measure the relative impact of its products and services on the member experience.
- The results are used for resource allocation and training of employees across the company.
- For this year we are planning to combine these models into one big model. Similar methodological issues will have to be addressed.

# Conclusions and recommendations

- Presented today are two means of dealing with situations where standard OLS regression would be problematic in partitioning out the relative impact of drivers.
- Multicollinearity was handled through Shapley Value Method.
  - Strong linkage to conceptual relationship of drivers to variance explained.
  - Stable estimates in the face of multicollinearity.
- Adjusted bivariate regressions were used to get reliable estimates of relationship that could then be rescaled to account for the incidence of the key drivers.
  - Allows for understanding relative importance when there is sparse data across predictors.