AAA Northeast Member Analysis

Yaroslav Klimenko

Contents

Background

AAA Northeast (AAA NE) is a regional clubs of the American Automobile Association (AAA).

- Core activity: **roadside assistance**, costly for AAA NE, particularly towing.
- Additional activities: other paid services - financial, insurance, travel products (aka **new products**).

AAA NE would like to minimize costs related to roadside assistance and maximize penetration (and thus revenues) of new products.

- Available data: AAA NE dataset with information on roadside assistance usage by members as well as other data on them.

Objectives

Provide a market segmentation of AAA households that allows AAA Northeast to better serve their members and optimize its financials.

Purpose of analysis:

- Better anticipate the needs of members
- Customize communications and offering to various segments
- Expend more effort driving acquisition and renewal of desirable members

Expectations from analysis:

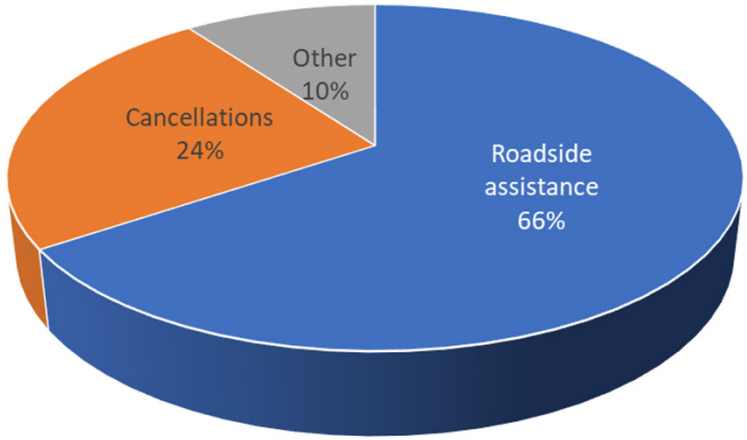
- Provide segmentation that enables to identify two types of members:
  - Those likely to yield higher roadside costs
  - Those likely to bring additional revenue through purchasing of new products

Approach

- Obtained data set and performed initial analysis
- Came back to client to clarify the data structure and the meaning of some variables
- Transformed data to get household based set
- Transformed variables where needed and identified variables of interest
- Created predictive models
  - Logistic regression (logit)
    - Likelihood of purchasing new products
    - Likelihood of roadside assistance usage
  - Linear regression
    - Costs of roadside assistance
- Built predictions
- Built customer segmentation using k-means and interpreted it

Original dataset

- Initial dataset
  - ~21,000 records from 2014 to 2019
  - 113 variables
- Mostly records related to roadside assistance or membership cancellations
- Many data missing
  - Data related to roadside assistance available only for records related to it
  - Some external data is not available for all members

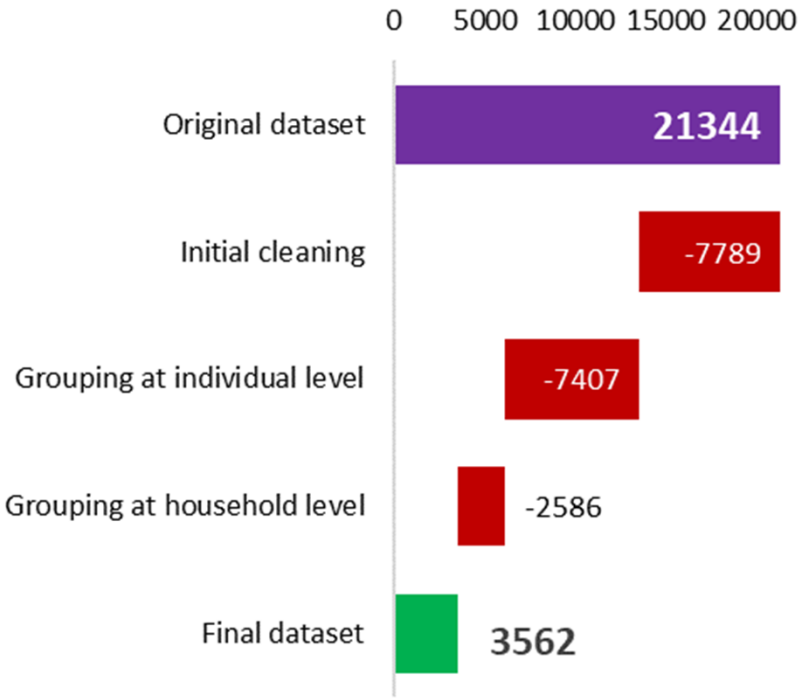


Data set composition

Data Preparation

- Data cleaning:
  - Remove outliers
    - 1 observation with almost all data (95 variables) missing
    - ~300 observations for the same household that seems to be a corporate customer)
  - Transform categorical variables into binary
  - Address missing values' issue:
    - Create additional variables for missing values (in most cases)

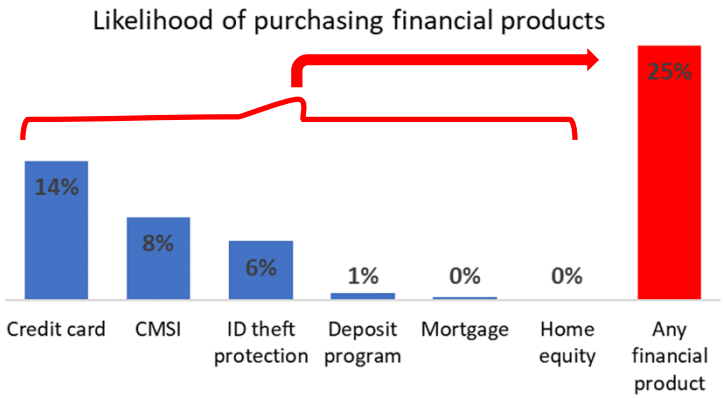
    or
    - fill missing values with appropriate values (mainly zeros)
  - Remove individuals that were not members in October 2018 or October 2019
- Aggregate data at individual level
  - Identify primary members and collect their attributes
- Aggregate data at household level
  - Create variables for number of household members in 2018 and 2019*



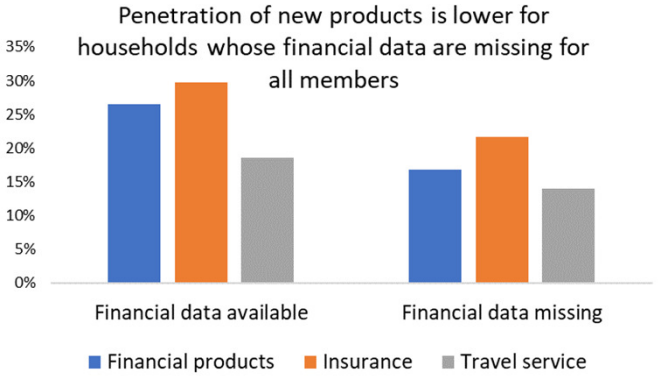
* Data were collected around 20 October 2019. For this reason, 2019 means precisely the period the last year before data collection, i.e. between October 2018 and October 2019. 2018 stands for a period between October 2017 and October 2018.
This notation is valid for the whole study.

Exploratory data analysis

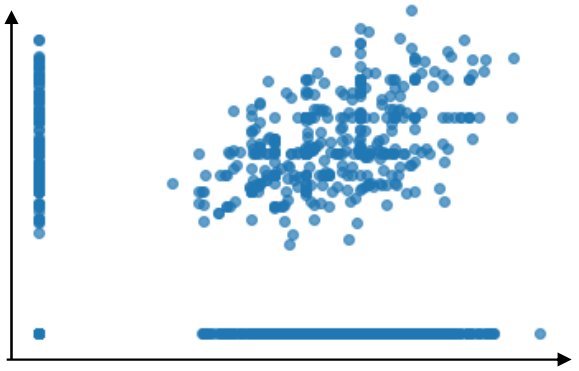
3 key findings



1. Too low penetration for many financial products to run a separate analysis
   → Grouped together



2. Households with missing values for financial variables are less likely to purchase a new product
   → Information does not miss at random



3. Evidence of correlation between current and past roadside assistance only for costs per member in logarithms
   → This configuration of costs was kept for analysis

Methodology: choice of models

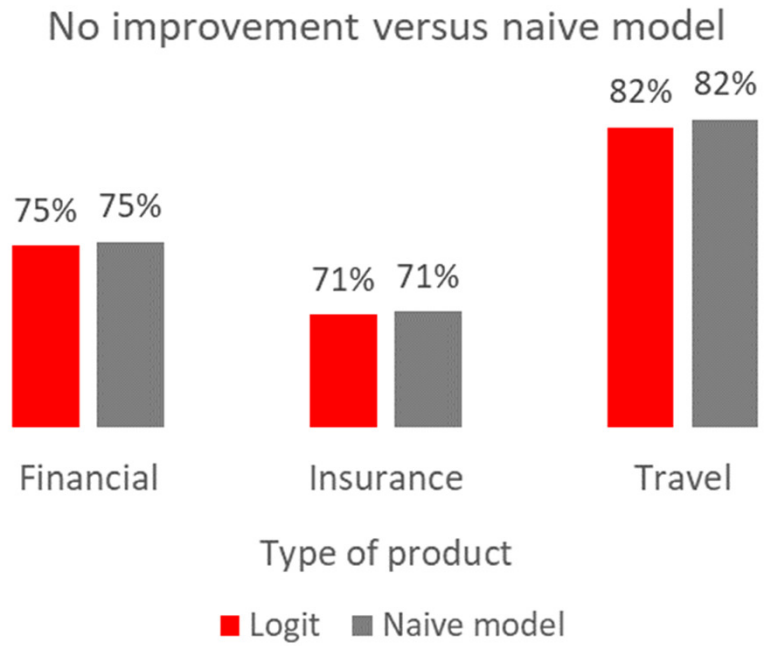
1. Evaluate likelihood of new product purchasing, roadside assistance usage and its costs:
   - Probability of roadside assistance usage or of new product purchasing: logistic regression
   - Costs of roadside assistance: linear regression
     - Evaluated on subsample of households that used roadside assistance in 2019 (i.e. costs conditional on roadside assistance usage)
     - Dependent variable : costs per household member in logarithms (the latter prevents from prediction of negative costs)
   - Evaluation is made on training subsample (75% of observations) and validated with testing subsample (25% of observations)
   - Explanatory variables (total 54 variables) :
     - Costs of roadside assistance in 2018
     - Number of household members in 2018
     - Socio-demographic and financial info on primary member
     - AAA NE Marketing related data
     - etc.
2. Compute probabilities of new product purchasing, roadside assistance usage and expected roadside assistance costs
   - For expected costs, predicted costs were multiplied by probability of roadside assistance usage:

     $E(costs) = E(costs \mid usage) * Pr(usage)$

   - For new products, replace probabilities with 1 where a purchase took actually place
3. Perform clustering of households using k-means
   - Interpret results

Analysis results: new products (1)

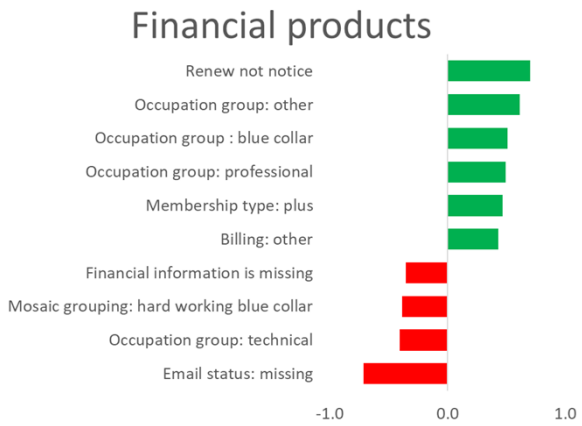
- 3 look-alike models for different target variables:
  - One or more of financial products
  - Insurance
  - Travel service
- Results:
  - Low predictive power in terms of predicted outcomes:
    - No improvement versus naive model*
  - It is a minor issue since only probabilities matter for further analysis
    - Intuition: identify users that have not purchased the products but are likely to purchase them

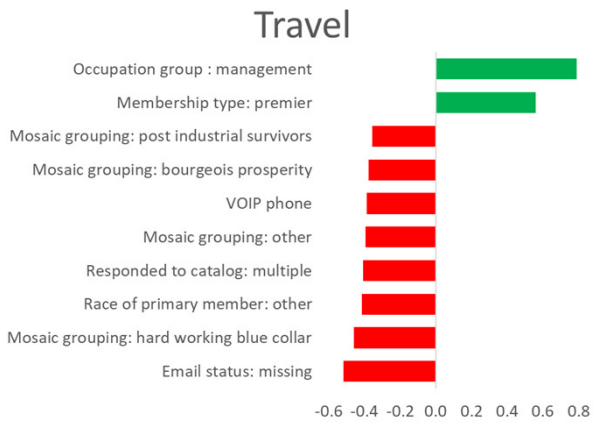


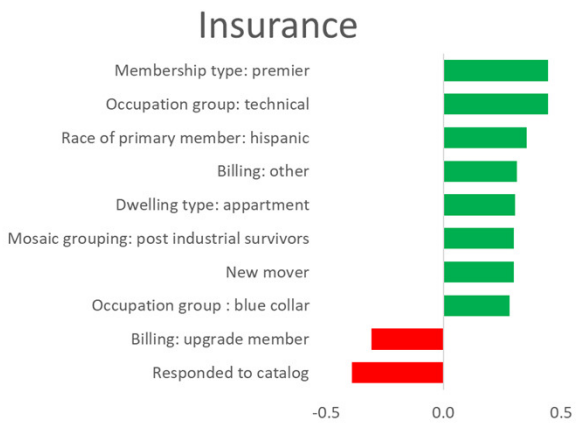
* Naive model: always predict the most common outcome (no product purchase)

Analysis results: new products (2)

Most important variable coefficients



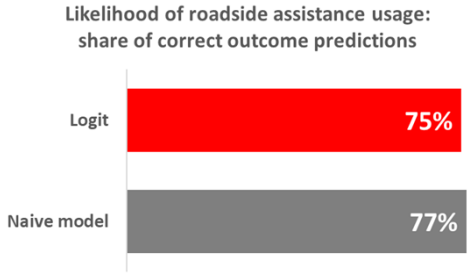


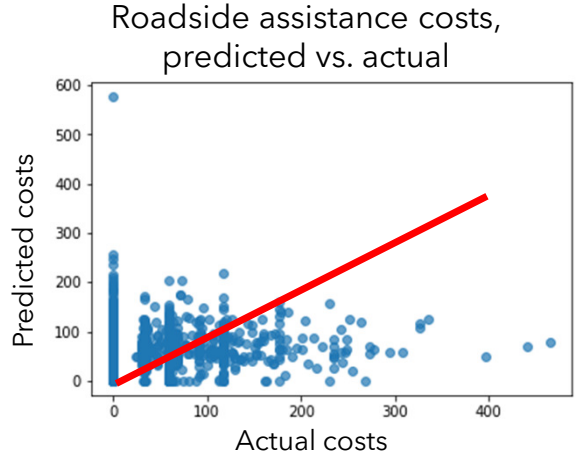


Membership type, occupation group of the primary member, mosaic household grouping and some administrative data (type of billing or phone) are important predictors of product purchase – as well as the fact of having some data missing

Analysis results: roadside assistance (1)

- Two models:
  - Logistic regression for probability to use roadside assistance, Pr(assistance)
  - Linear regression for costs, performed on the subsample of households that used roadside assistance in 2019 (costs conditioning on assistance, costs | assistance)
- Low predictive power for both models:
  - For probability of roadside assistance, no improvement versus naive case (predict no roadside assistance for all households)
  - ~25% of variance in roadside assistance costs are explained with linear regression, with model failing to predict correctly high costs

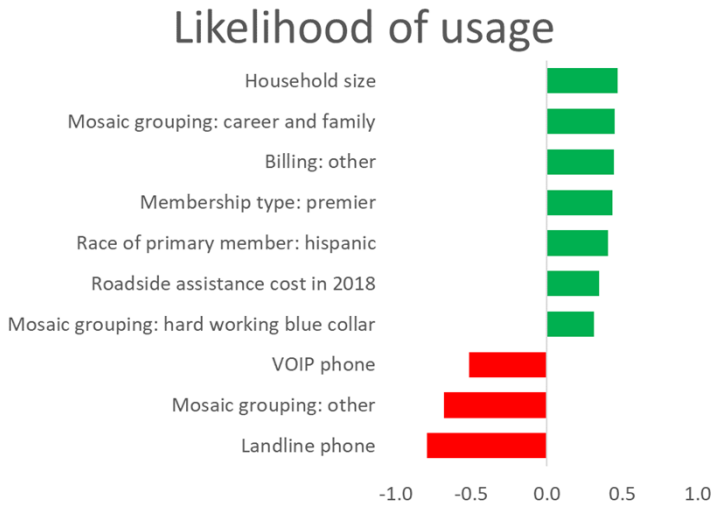


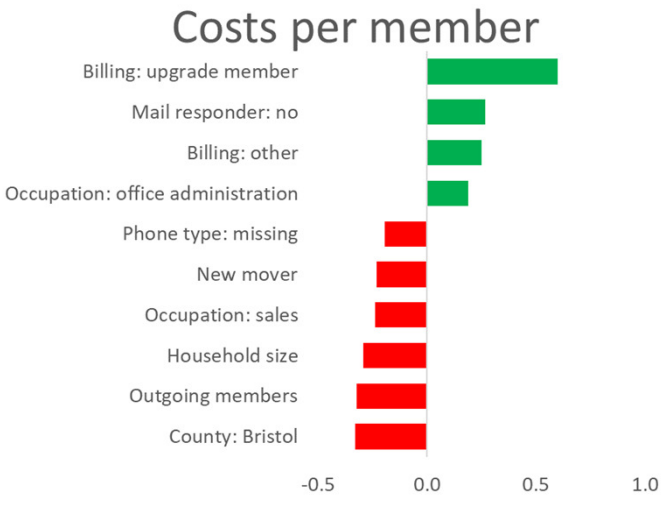


* Along the red line, predictions are equal to actuals.
For points below the line, actual is higher than prediction.

Analysis results: roadside assistance (2)

Most important variable coefficients





- Likelihood of usage is influenced by household size, mosaic grouping, membership type and previous year's roadside assistance costs.
- Costs of roadside assitance depend mainly on household size and the occupation of the primary member.
- Previous year's costs increase likelihood of using roadside assistance in the current year but do no have impact on current year's costs. Correlation shown in EDA part is not cofirmed by regression analysis.

Analysis results: clustering (1)

- Data preparation:
  - Probabilities of new product purchasing, roadside assistance usage and expected roadside assistance costs were computed
  - For expected costs, predicted costs were multiplied by probability of roadside assistance usage:

    $$E(costs) = E(costs \mid usage) * Pr(usage)$$

  - For new products, replace probabilities with 1 where a purchase took actually place
- K-Means clustering using these variables
  - Only probabilities were used, without expected costs. Adding the latter did not allow a clear split into clusters.
- Determine number of clusters:
  - Decompose the variables into 2 dimensions using principal component analysis
  - Visualize data
  - 5 clusters kept for the final analysis

Visualization of customers' clusters in 2-dimensional space

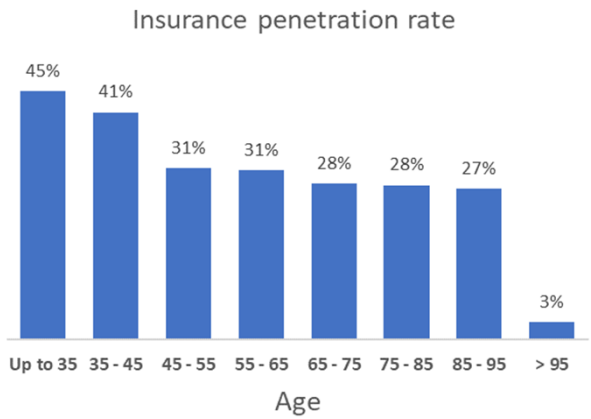
Analysis results: clustering (2)

- Clusters interpretation:
  1. Active AAA clients (10% of sample)
     Relatively young users with high revenue and credit ratings, frequently occupied in management or administration. Relatively often are renters (big cities' residents?)
     They would be AAA NE's best friends if they weren't generating biggest roadside assistance costs
  2. Users of financial products (12%)
     Similar to the first cluster but with roadside assistance costs at the average level. Many of them are retired. Other characteristics are average.
  3. Users of travel products (15%)
     Average roadside assistance costs, usage of travel products and moderate usage of financial services. They have highest income among all clusters and often have management positions.
  4. Users of insurance products (16%)
     Average roadside assistance costs, usage of insurance products, and almost no usage of other new products. Often retired.
  5. Unexplored territory (47%)
     Do not use any new product, have roadside assistance costs below average, Many of 'Bourgeous prosperity', as classified by Mosaic classification are there, and relatively few of them have 'Plus' membership. Most members with missing financial data are in this cluster. Age is above average. Taking into account low data available and low activity, some of them may already be dead.

Clusters	1	2	3	4	5
Size of cluster	341	428	549	567	1677
Part of sample	10%	12%	15%	16%	47%
Purchased Financial product	97%	98%	25%	0%	0%
Purchased Insurance	97%	0%	21%	99%	0%
Purchased Travel product	25%	0%	100%	0%	0%
Roadside assistance costs	49	38	41	43	34
Plus membership	38%	35%	37%	31%	31%
fin_missing	7%	15%	14%	15%	21%
Number of members in household	2.11	1.76	1.86	1.57	1.44
Average age	67	70	70	71	73
Average children number	0.93	0.64	0.71	0.66	0.53
Primary member: male	65%	58%	60%	59%	50%
Credit ranges (the more the better)	4.8	4.5	4.5	4.4	4.0
Yearly income, k-USD	90	88	95	77	77
Dwelling type: multi-family with appt number	11%	8%	9%	13%	11%
Home renter	6%	2%	3%	6%	4%
Mosaic: bourgeois prosperity	14%	15%	17%	15%	18%
Mosaic: routine service workers	15%	12%	13%	11%	9%
Occupation: management	11%	7%	11%	6%	5%
Occupation: administration	11%	4%	4%	3%	2%
Occupation: professional	23%	16%	16%	10%	11%
Occupation: retired	22%	28%	25%	30%	24%

Recommendations

- Recommendations & Conclusions:
  - No good or bad profiles in terms of road assistance costs: year-on-year correlation between costs is relatively low, today's costs do not impact much tomorrow's costs.
  - Buyers of Travel products are a potential target for insurance and financial products (Cluster 3)
  - Get to know your those whom you don't know today (Cluster 5)
  - Check whether all AAA members are alive ☺ (Cluster 5)
  - Attract new young members
    - Between 2018 and 2019, only 70 new members joined while ~300 left
    - Insurance products are most attractive among relatively young population



Next steps

- Next steps:
  - Clarify the data construction with AAA Northeast
    - Many things are still unclear
    - More data should be available internally
  - Try other models to improve predictions:
    - Random forest
    - Lasso regression for a better feature selection
    - More adapted models for costs (tobit or a more complex ad-hoc modelling)
    - Different clustering method (Markov models or different distance measure)

Appendix

Original dataset (2)

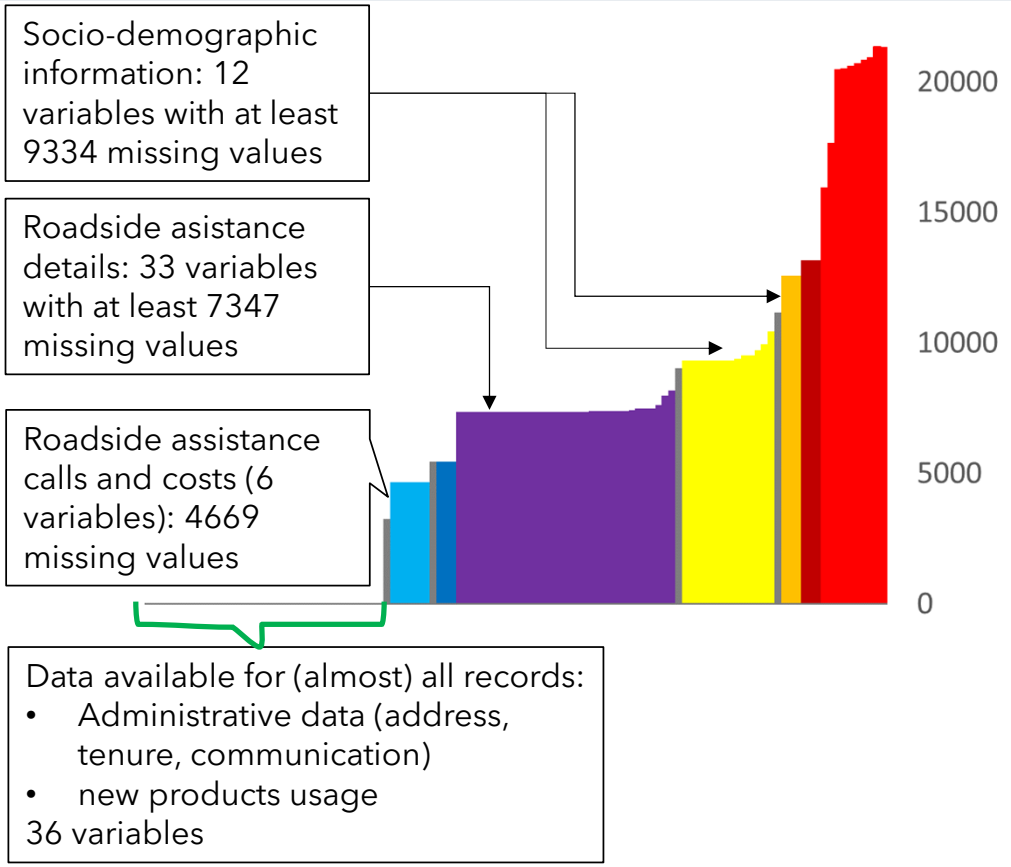
- AAA NE Enriches internal data set with external information:
  - Socio-demographic data
  - Credit and financial data

→ Many missing data:
  - If record not related to roadside assistance
  - If external data not available

→ No observations without missing data

→ Data does not miss at random

- In addition, data for period between mid 2017 and mid 2018 are missing, which limited seriously the possibilities of analysis and made available data less precise



List of explanatory variables

Category	Variable
Household	Age of the primary member at the beginning of the current period
	Household size at the beginning of the current period
	Outgoing members: number of members having left the household in last 365 days**
Membership	Came to AAA NE from another AAA club: no*
	Came to AAA NE from another AAA club: yes
	Membership type: Basic*
	Membership type: Plus
	Membership type: Premier
	Motorcycle indicator: missing
	Motorcycle indicator: no*
	Motorcycle indicator: yes
	Roadside assistance costs in 2018**
	Year of club joining**
Marketing	Billing: Renewal Mbr billing*
	Billing: Upgrade member
	Billing: Other
	Email status: active*
	Email status: unsubscribed
	Email status: other
	Email status: missing
	Mail responder: yes*
	Mail responder: no
	Member phone type: wireless*
	Member phone type: landline
	Member phone type: VOIP
	Member phone type: unknown or missing
	Renew method: notice*
	Renew method: no notice (other)
	Responded to catalog: no*
	Responded to catalog: yes
	Responded to catalog: multiple responses

Category	Variable
Socio-demographic	County: Bristol
	County: Kent
	County: Newport
	County: Providence or other*
	County: Washington
	Credit range of primary member**
	Dwelling type: apartment
	Dwelling type: SFDU or PO box*
	Education of the primary member: completed college
	Education of the primary member: high school
	Education of the primary member: some college or graduated school*
	Education of the primary member: unknown
	Financial info missing
	Gender of primary member: female*
	Gender of primary member: male
	Gender of primary member: missing
	Income of primary member**
	Language of primary member: known*
	Language of primary member: unknown
	Mosaic household grouping: Bourgeois Prosperity
	Mosaic household grouping: Career and Family
	Mosaic household grouping: Comfortable Retirement*
	Mosaic household grouping: Hard Working Blue Collar
	Mosaic household grouping: missing
	Mosaic household grouping: Other
	Mosaic household grouping: Post Industrial Survivors
	Mosaic household grouping: Routine Service Workers
	New mover: no*
	New mover: yes
	Number of children**
	Occupation group of primary member: blue collar
	Occupation group of primary member: management
	Occupation group of primary member: office administration
	Occupation group of primary member: other
	Occupation group of primary member: professional
	Occupation group of primary member: retired*
	Occupation group of primary member: sales
	Occupation group of primary member: technical
	Race of primary member: European / Caucasian*
	Race of primary member: hispanic
	Race of primary member: other

* Reference category
** Numeric variable, used in models in standardized form

Both variables related to motorcycle ownership were used only in models for new products. This info is available only if household called roadside assistance, so usage of this variable to predict roadside assistance or its costs would bias results

Flowchart to establish the member status in Oct 2019

To calculate the number of members in household in Oct 2018 and Oct 2019, it was crucial to understand which members were active at that time. The flowchart shown here describes decision process.

Similar flowcharts were applied for other variables (e.g. to date all records)