Caravan Insurance Policy

Naveen Vijayakumar

AGENDA

01 Introduction

02 Dataset Overview

03 Business Objectives

04 Data Pre-processing

05 Data Mining Tasks and Techniques

- Prediction
- Cross-Selling
- Identifying High Value Customers

06 Conclusion

Caravan Insurance Policy - Insurance Company Benchmark [CoIL (Computational intelligence and Learning) 2000]

Link - https://kdd.ics.uci.edu/databases/tic/tic.data.html

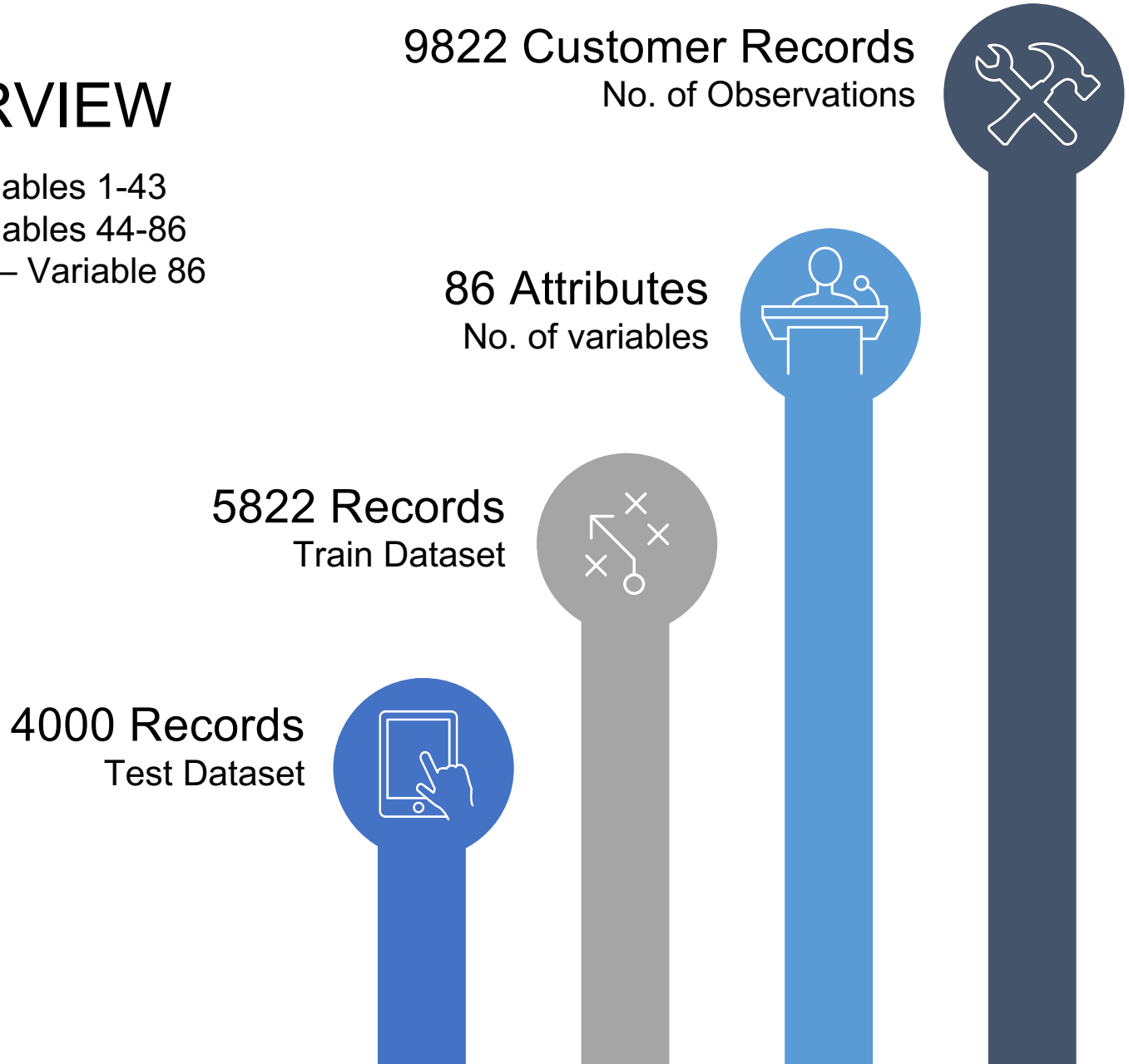
INTRODUCTION

Caravan Insurance Policy

This data set used in the 2000 edition of the COIL Challenge contains information on customers of an insurance company. The data consists of 86 variables and includes product usage (policy ownership) data and socio-demographic data.

DATASET OVERVIEW

Sociodemographic data - Variables 1-43
Product ownership data - Variables 44-86
Caravan insurance purchase – Variable 86



DATASET OVERVIEW

Before

33	1	3	2	8	0	5	1	3	7
2	5	2	1	1	2	6	1	1	8
3	0	0	0	6	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0								
37	1	2	2	8	1	4	1	4	6
5	0	4	0	2	3	5	0	2	7
4	2	0	0	0	0	0	0	0	0
0	2	0	0	0	0	0	0	0	0
0	0								
37	1	2	2	8	0	4	2	4	3
7	0	2	0	5	0	4	0	7	2
4	2	0	0	6	0	0	0	0	0
0	1	0	0	1	0	0	0	0	0
0	0								
9	1	3	3	3	2	3	2	4	5
3	1	2	3	2	1	4	0	5	4
4	0	0	0	6	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0								
40	1	4	2	10	1	4	1	4	7
0	0	0	9	0	0	0	0	4	5
3	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0								
23	1	2	1	5	0	5	0	5	0
4	2	2	2	2	2	4	2	9	0
3	0	0	0	6	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0								
39	2	3	2	9	2	2	0	5	7
4	1	5	0	1	4	5	0	6	3
5	0	0	0	6	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0

After

ORIGIN	MOSTYPE	MAANTHUI	MGEMOMV	MGEMLEEF	MOSHOOFD	MGODRK	MGODPR	MGODOV	MGODGE
train	33	1	3	2	8	0	5	1	3
train	37	1	2	2	8	1	4	1	4
train	37	1	2	2	8	0	4	2	4
train	9	1	3	3	3	2	3	2	4
train	40	1	4	2	10	1	4	1	4
train	23	1	2	1	5	0	5	0	5
train	39	2	3	2	9	2	2	0	5
train	33	1	2	3	8	0	7	0	2
train	33	1	2	4	8	0	1	3	6
train	11	2	3	3	3	3	5	0	2
train	10	1	4	3	3	1	4	1	4
train	9	1	3	3	3	1	3	2	4
train	33	1	2	3	8	1	4	1	4
train	41	1	3	3	10	0	5	0	4
train	23	1	1	2	5	0	6	1	2
train	33	1	2	3	8	0	7	0	2
train	38	1	2	3	9	0	6	0	3
train	22	2	3	3	5	0	5	0	4
train	13	1	4	2	3	2	4	0	3
train	31	1	2	4	7	0	2	0	7
train	33	1	4	3	8	0	6	0	3
train	33	2	3	3	8	0	4	2	3
train	13	1	3	2	3	1	7	0	2
train	34	2	3	2	8	0	7	0	2
train	13	2	4	3	3	0	4	2	4
train	33	1	3	3	8	0	6	1	2
train	37	1	3	3	8	0	5	0	4
train	40	1	3	3	10	0	3	0	6
train	31	1	4	2	7	0	9	0	0
train	33	2	2	3	8	0	7	1	2

DATASET OVERVIEW

Socio-demographic data

1 MOSTYPE Customer Subtype see L0
2 MAANTHUI Number of houses 1 ñ 10
3 MGEMOMV Avg size household 1 ñ 6
4 MGEMLEEF Avg age see L1
5 MOSHOOFD Customer main type see L2
6 MGODRK Roman catholic see L3
7 MGODPR Protestant ...
8 MGODOV Other religion
9 MGODGE No religion
10 MRELGE Married

Product ownership data

44 PWAPART Contribution private third party insurance see L4
45 PWABEDR Contribution third party insurance (firms) ...
46 PWALAND Contribution third party insurane (agriculture)
47 PPERSAUT Contribution car policies
48 PBESAUT Contribution delivery van policies
49 PMOTSCO Contribution motorcycle/scooter policies
50 PVRAAUT Contribution lorry policies
51 PAANHANG Contribution trailer policies
52 PTRACTOR Contribution tractor policies
53 PWERKT Contribution agricultural machines policies
54 PBROM Contribution moped policies

L0:

Value Label
1 High Income, expensive child
2 Very Important Provincials
3 High status seniors
4 Affluent senior apartments
5 Mixed seniors
6 Career and childcare
7 Dinki's (double income no kids)
8 Middle class families
9 Modern, complete families
10 Stable family

L1:

1 20-30 years
2 30-40 years
3 40-50 years
4 50-60 years
5 60-70 years
6 70-80 years

L2:

1 Successful hedonists
2 Driven Growers
3 Average Family
4 Career Loners
5 Living well
6 Cruising Seniors
7 Retired and Religeous
8 Family with grown ups
9 Conservative families
10 Farmers

L3:

0 0%
1 1 - 10%
2 11 - 23%
3 24 - 36%
4 37 - 49%
5 50 - 62%
6 63 - 75%
7 76 - 88%
8 89 - 99%
9 100%

L4:

0 f 0
1 f 1 ñ 49
2 f 50 ñ 99
3 f 100 ñ 199
4 f 200 ñ 499
5 f 500 ñ 999
6 f 1000 ñ 4999
7 f 5000 ñ 9999
8 f 10.000 - 19.999
9 f 20.000 - ?

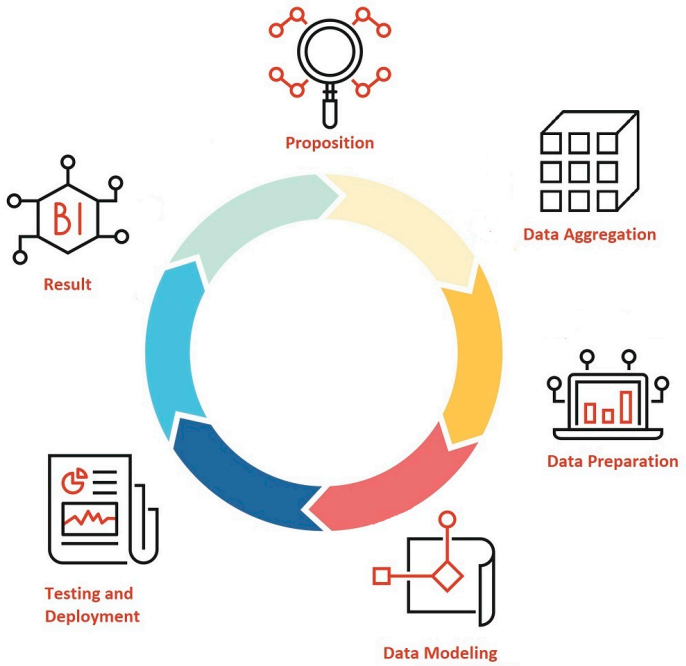
BUSINESS OBJECTIVES

1. Predict customers who will buy Caravan Insurance based on their socio-demographic data and their purchase behavior

2. Identify opportunities to cross-sell using association rule mining

3. Identify high value customers subtype for effective targeting

PREPROCESSING

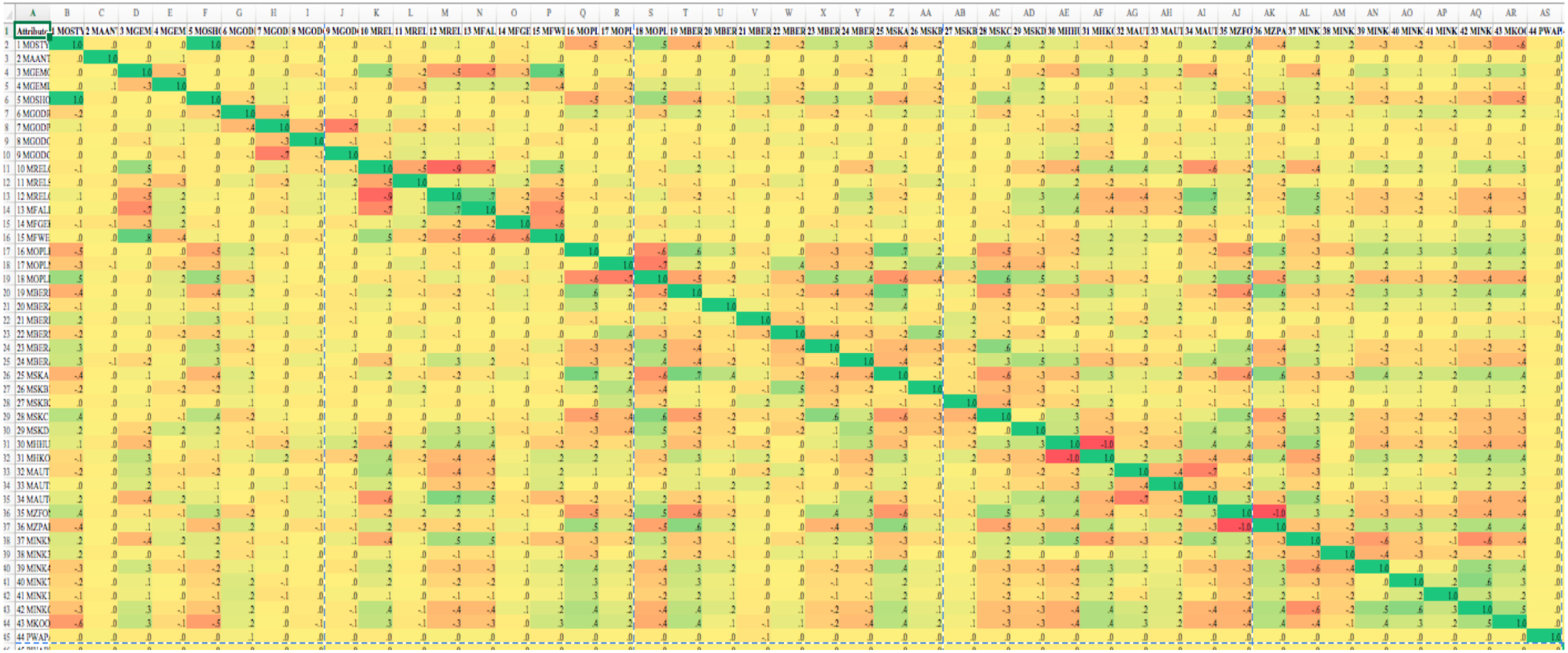


- Noise
- Outliers
- Duplicates
- Missing values
- 86 attributes with customer demographics & psychographics data
- 9822 records (total): 5822 train data set and 4000 test records

PREPROCESSING

- Understanding attributes through data assessment
- Transformed the variables in ranges to an approximate value
- Train and predicted variable CARAVAN was transformed to binomial
- Select attributes using correlation matrix

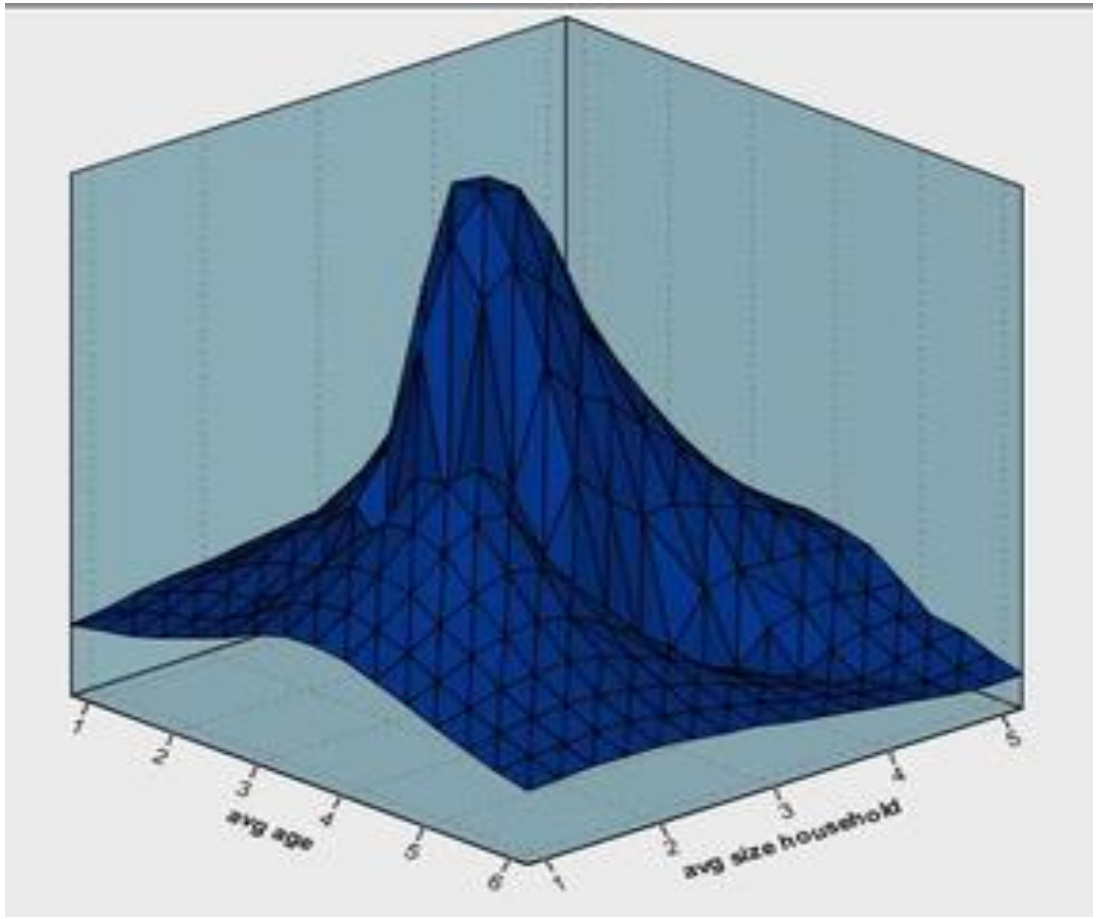
PREPROCESSING- SELECTING ATTRIBUTES



SELECTED ATTRIBUTES

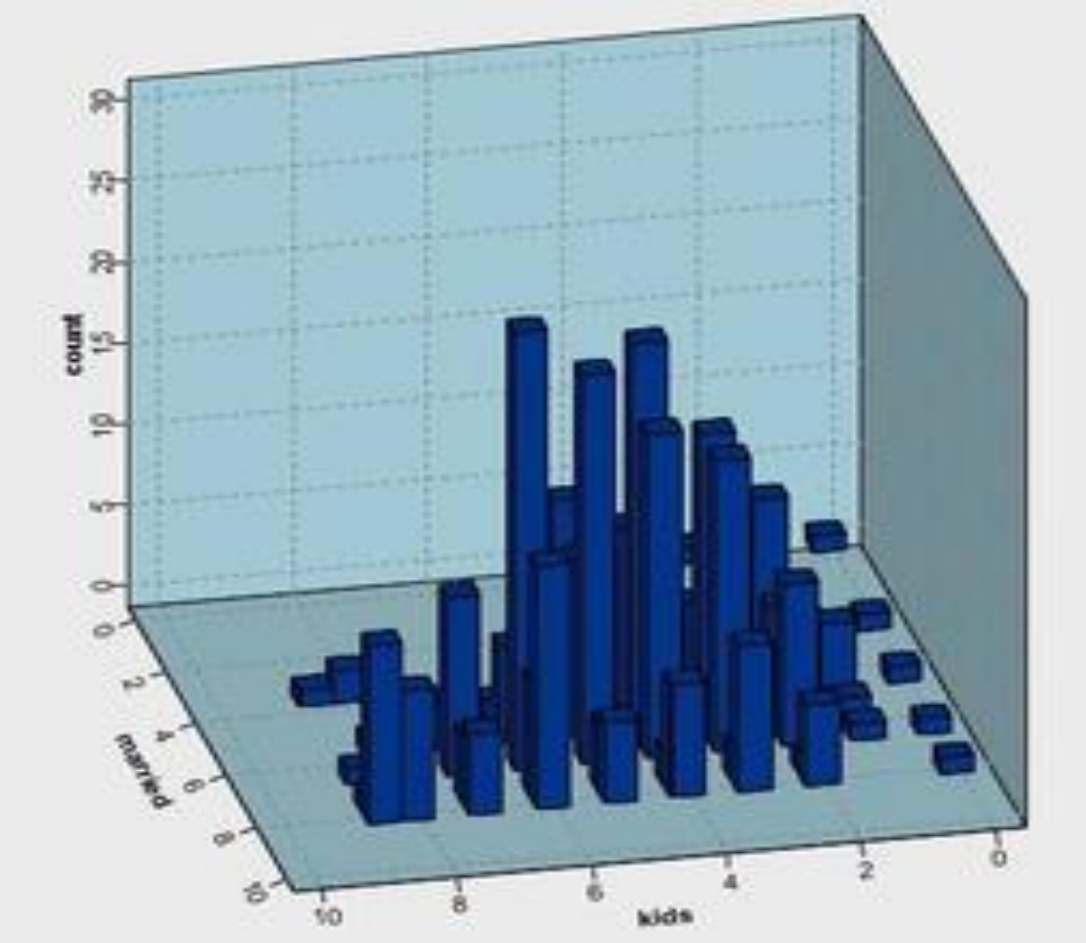
1	ABROM - Number of moped policies	9	PAANHANG - Contribution trailer policies
2	APLEZIER - Number of boat policies	10	PBRAND - Contribution fire policies
3	CARAVAN - Number of mobile home policies 0 - 1	11	PLEVEN - Contribution life insurances
4	MINK4575 - Income 45-75.000	12	PPERSAUT - Contribution car policies
5	MINKM30 -Income < 30.000	13	PPLEZIER - Contribution boat policies
6	MBERMIDD-Middle Management – Employment data	14	MHHUUR Rented house
7	MAUT0 - No car	15	MFALLEEN Singles
8	MGODGE- No religion	16	MSKD Social Class D

INITIAL HYPOTHESIS



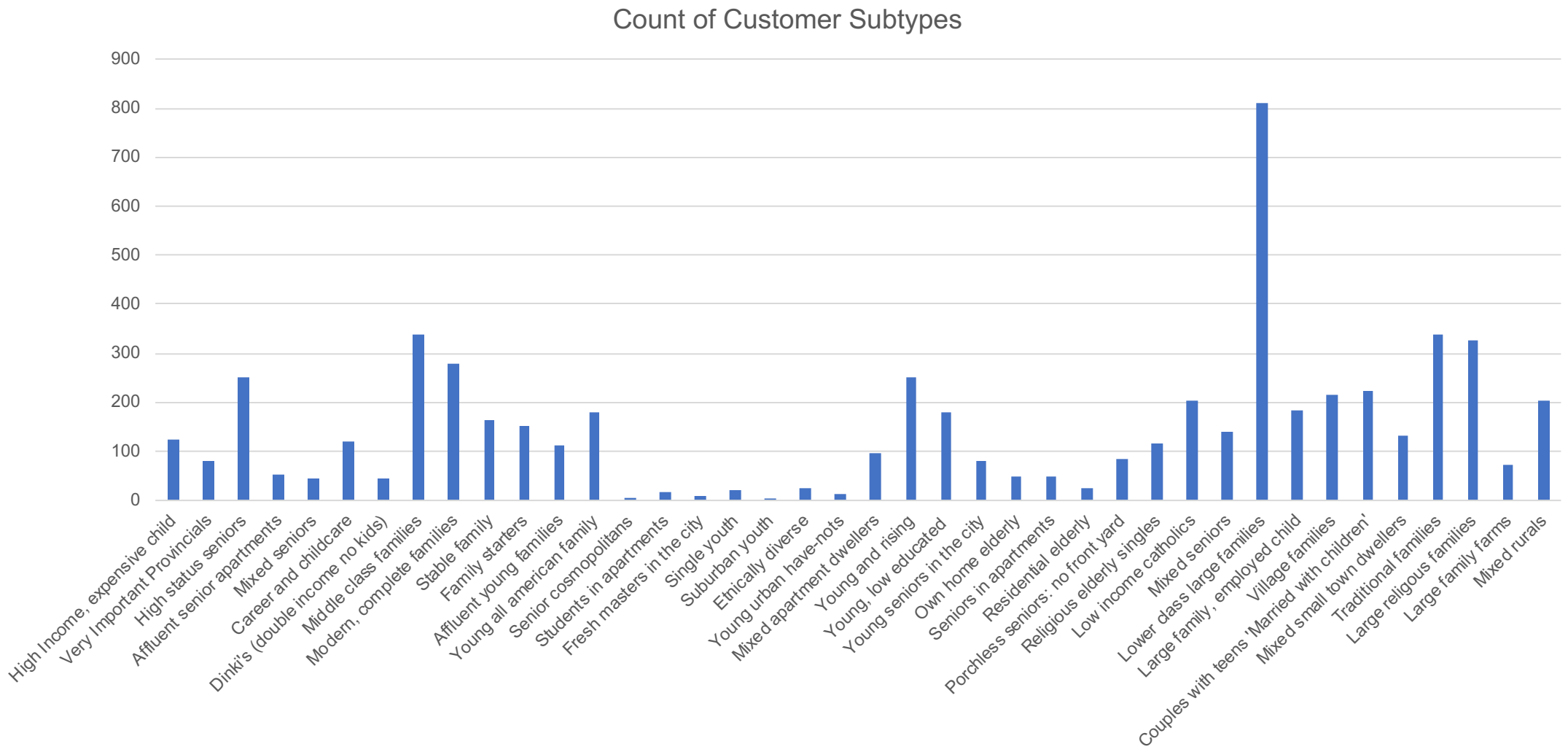
- Majority of caravan customers are in the average age group of 2 which is between 30-40 year olds with average household size of 4 members

INITIAL HYPOTHESIS



- Data points of both ’Married’ and ‘Kids’ are in the level range of 5
- From which we can conclude that the customers have an equal chance of being married and having children

INITIAL HYPOTHESIS



INTIAL HYPOTHESIS

Few hypothesis from the above visualization:

1. Caravan customers are of two groups:
   - Middle class families
   - Lower class large families
2. Targeted caravan customers are of the average age group 30-40
3. Customers who buy boat insurance are likely to buy caravan insurance
4. Other products that caravan insurance customers buy are car and file insurance policies

PREDICTION- COMPARISON OF METHODS

- Cross Validation
  - Naïve Bayes
  - Logistic Regression
  - Neural Net

- ROC Curves

PREDICTION- COMPARISON OF METHODS

Naive Bayes:

Accuracy : 92.12% + /- 0.70% (Mikro: 92.12%)	Column1	Column2	Column3
	True 0	True 1	Class Prediction
Pred. 0	5336	321	94.33%
pred. 1	138	27	16.36%
Class Recall	97.43%	7.76%	

PREDICTION- COMPARISON OF METHODS

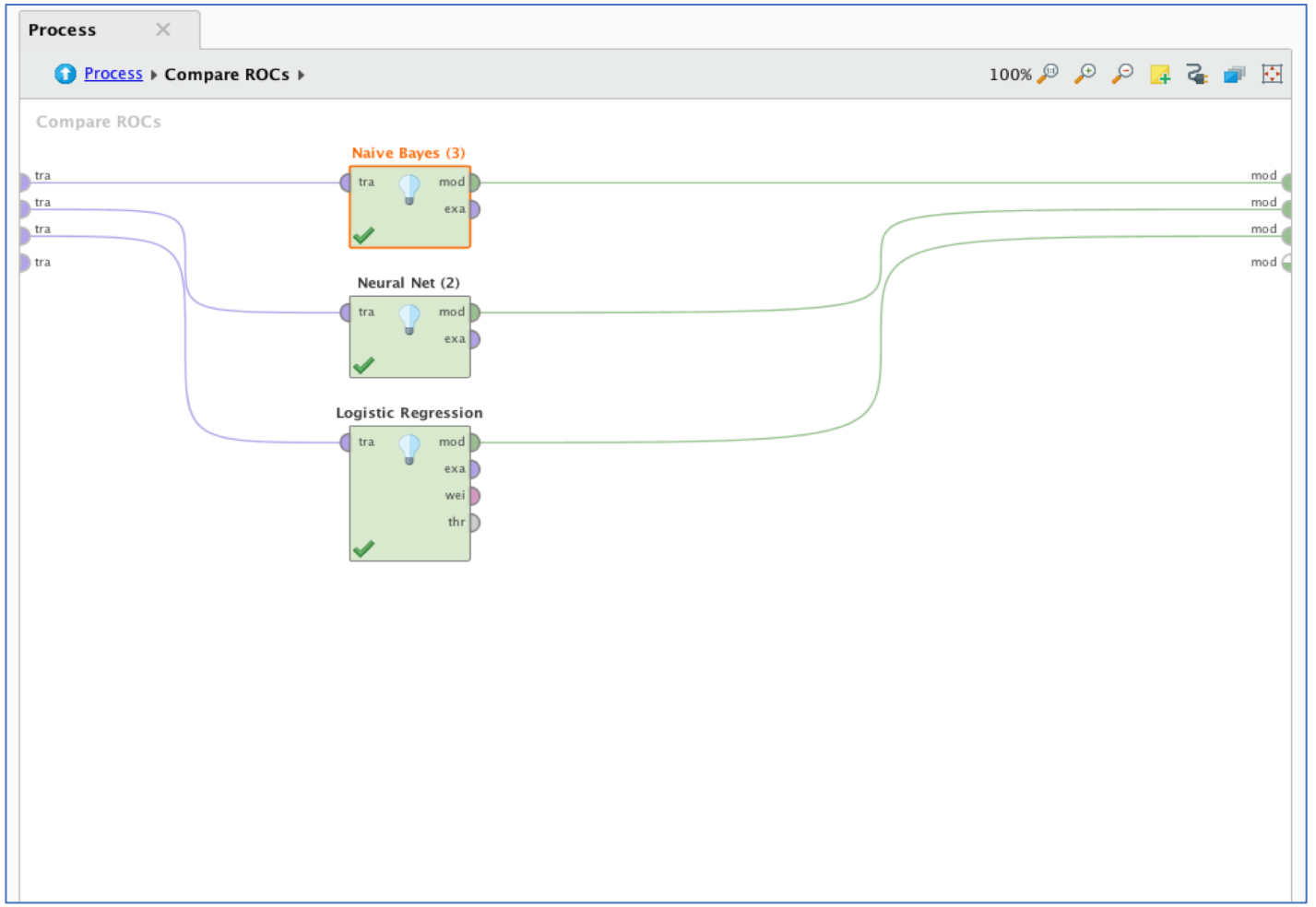
Logistic Regression:

Accuracy : 92.12% + /- 0.70% (Mikro: 92.12%)	Column1	Column2	Column3
	True 0	True 1	Class Prediction
Pred. 0	5336	321	94.33%
pred. 1	138	27	16.36%
Class Recall	97.43%	7.76%	

Accuracy : 92.12% + /- 0.70% (Mikro: 92.12%)	Column1	Column2	Column3
	True 0	True 1	Class Prediction
Pred. 0	5336	321	94.33%
pred. 1	138	27	16.36%
Class Recall	97.43%	7.76%	

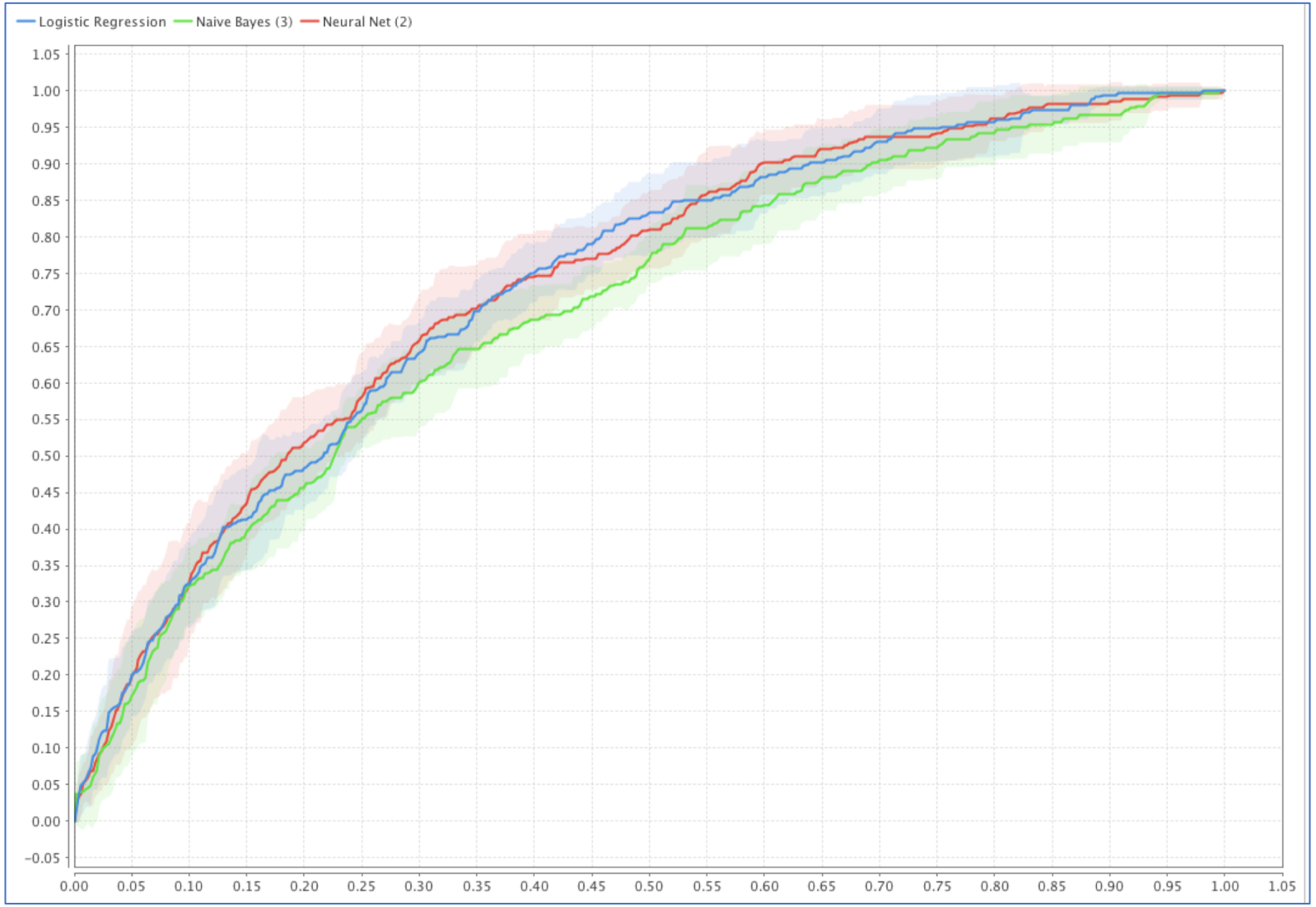
PREDICTION- COMPARISON OF METHODS

Rapid miner process for prediction task



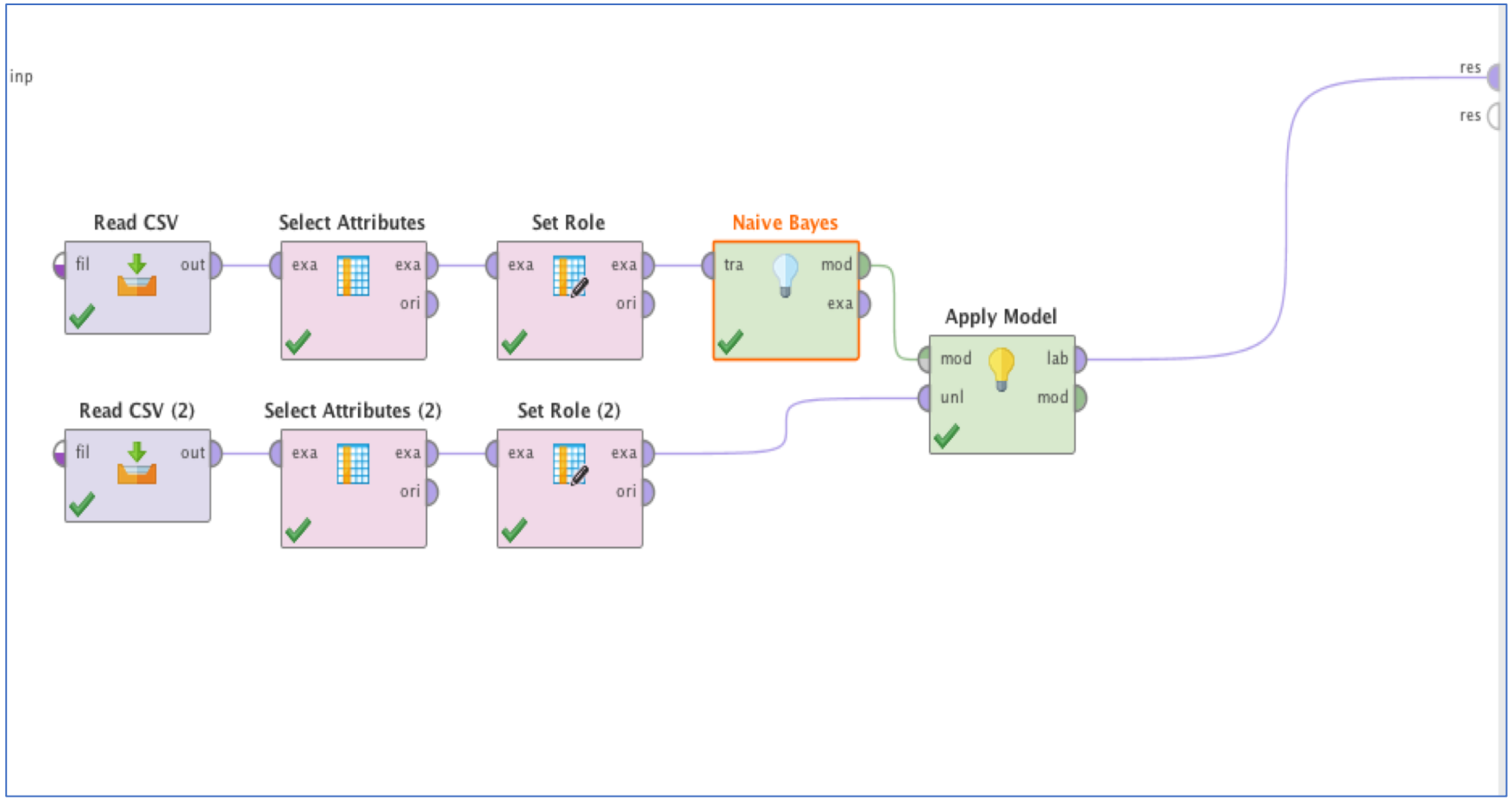
PREDICTION- COMPARISON OF METHODS

ROC Curve output

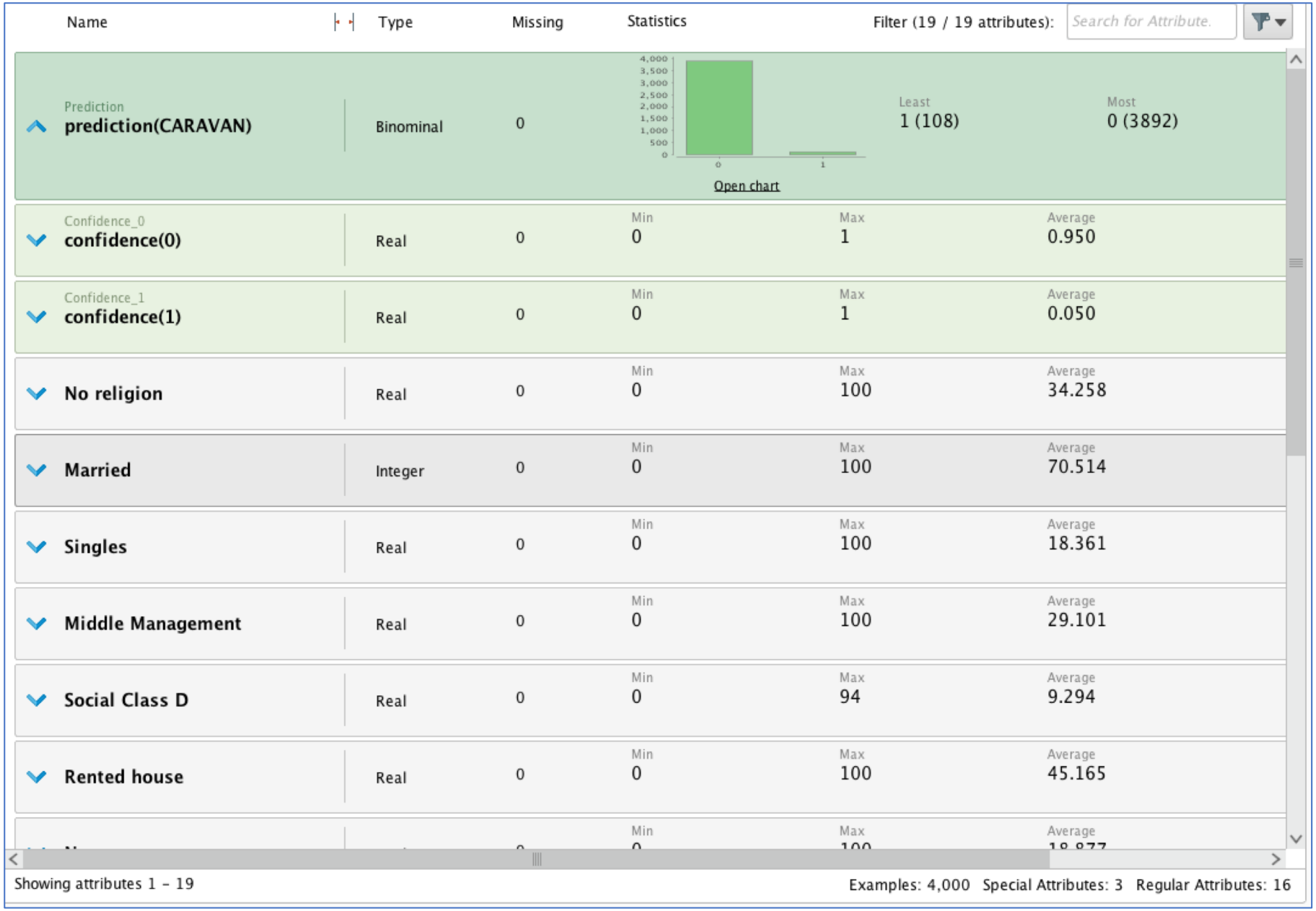


PREDICTION

Complete Rapid miner process flow layout



RESULT



Name	Type	Missing	Statistics		
Prediction **prediction(CARAVAN)**	Binominal	0	 Open chart	Least 1 (108)	Most 0 (3892)
Confidence_0 **confidence(0)**	Real	0	Min 0	Max 1	Average 0.950
Confidence_1 **confidence(1)**	Real	0	Min 0	Max 1	Average 0.050
No religion	Real	0	Min 0	Max 100	Average 34.258
Married	Integer	0	Min 0	Max 100	Average 70.514
Singles	Real	0	Min 0	Max 100	Average 18.361
Middle Management	Real	0	Min 0	Max 100	Average 29.101
Social Class D	Real	0	Min 0	Max 94	Average 9.294
Rented house	Real	0	Min 0	Max 100	Average 45.165

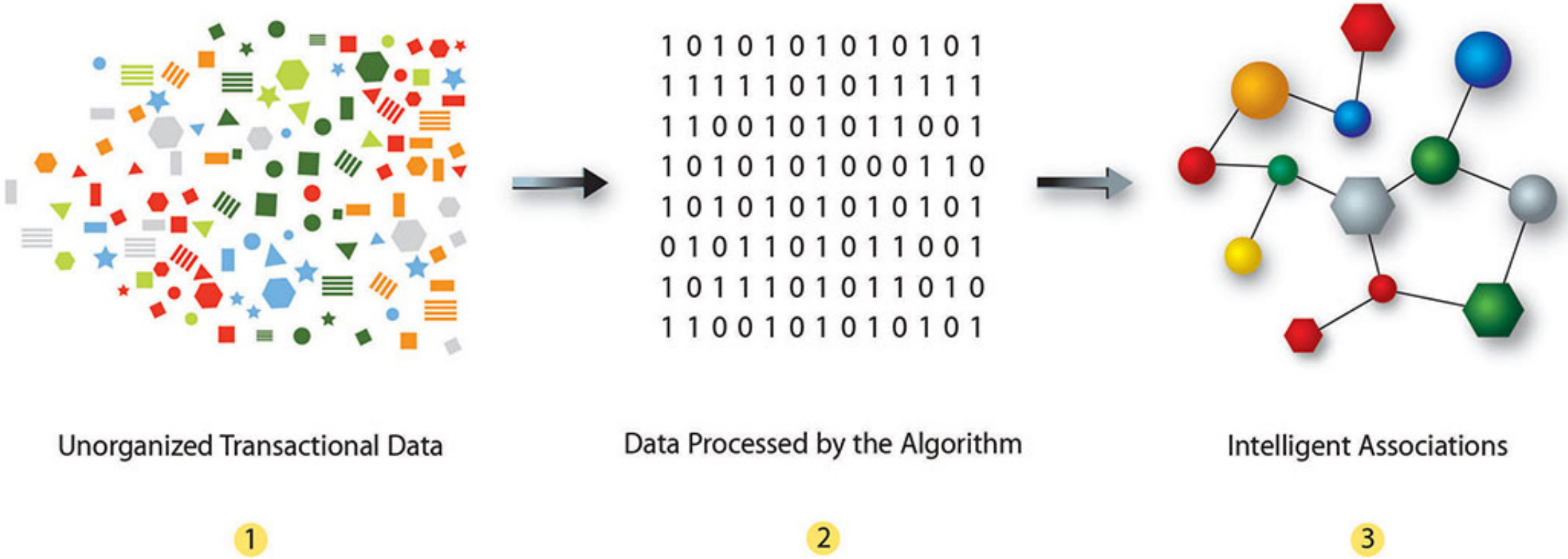
PREDICTION

Using Naive Bayes Classifier we predicted 108 customers to buy the Caravan Insurance.

Real World Application

With this model, we can help the insurance company to better target their customers. We can reduce their input marketing costs by refining their mailing lists with prospective customers.

ASSOCIATION RULE MINING (ARM)



Transformed 22 variables showing number of policies owned into Binary variables

Before

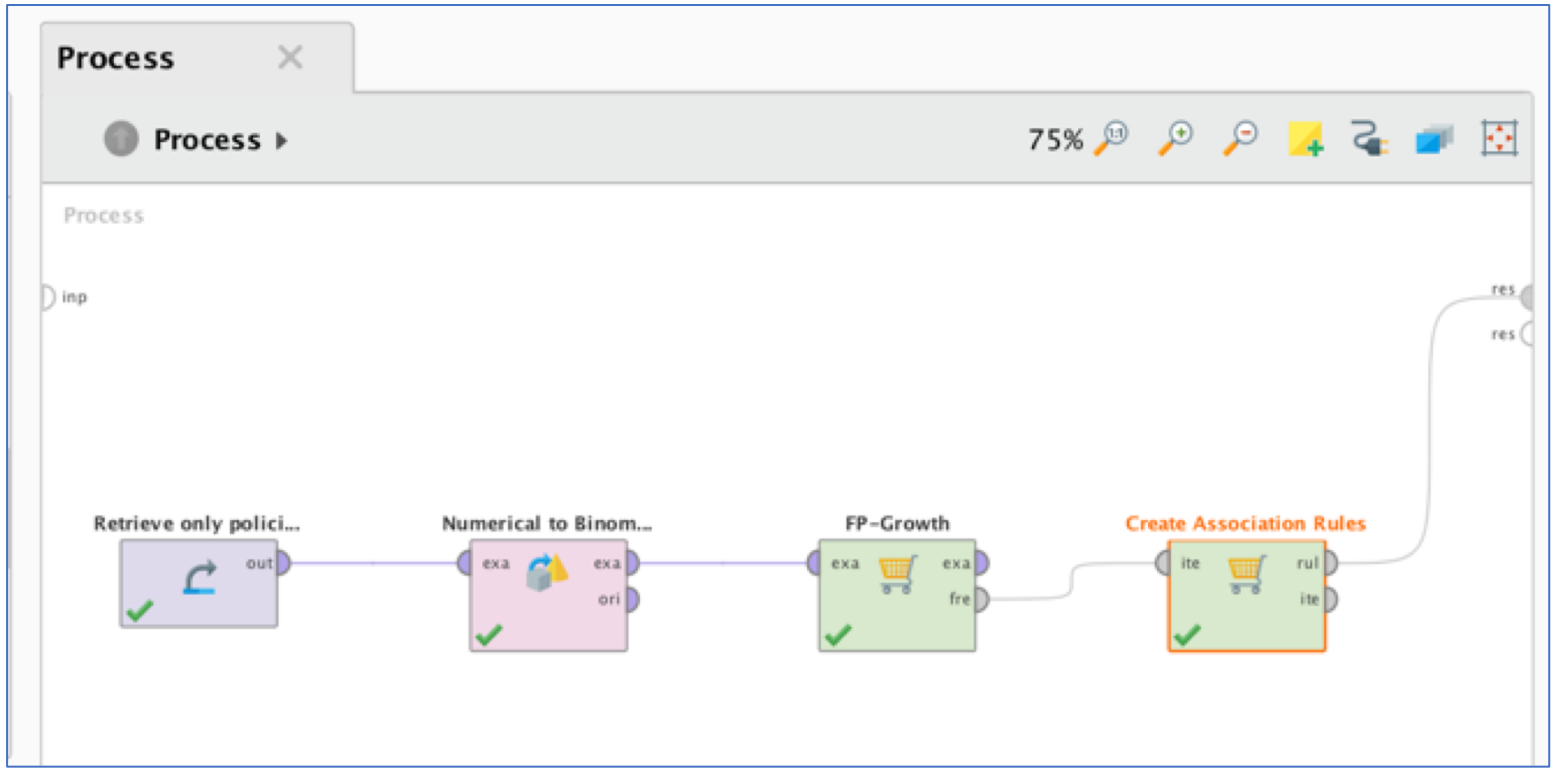
33	1	3	2	8	0	5	1	3	7
2	5	2	1	1	2	6	1	1	8
3	0	0	0	6	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0								
37	1	2	2	8	1	4	1	4	6
5	0	4	0	2	3	5	0	2	7
4	2	0	0	0	0	0	0	0	0
0	2	0	0	0	0	0	0	0	0
0	0								
37	1	2	2	8	0	4	2	4	3
7	0	2	0	5	0	4	0	7	2
4	2	0	0	6	0	0	0	0	0
0	1	0	0	1	0	0	0	0	0
0	0								
9	1	3	3	3	2	3	2	4	5
3	1	2	3	2	1	4	0	5	4
4	0	0	0	6	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0								
40	1	4	2	10	1	4	1	4	7
0	0	0	9	0	0	0	0	4	5
3	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0								
23	1	2	1	5	0	5	0	5	0
4	2	2	2	2	2	4	2	9	0
3	0	0	0	6	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0								
39	2	3	2	9	2	2	0	5	7
4	1	5	0	1	4	5	0	6	3
5	0	0	0	6	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0

After

Private third party insurance	Car policies	Delivery van policies	Motorcycle policies	Lorry policies
0	1	0	0	0
1	0	0	0	0
1	1	0	0	0
0	1	0	0	0
0	0	0	0	0
0	1	0	0	0
0	1	0	0	0
0	0	0	0	0
0	1	0	0	0
1	0	0	0	0
0	1	0	0	0
1	1	0	0	0
0	1	0	0	0
0	0	0	0	0
1	1	0	0	0
1	0	0	0	0
0	0	0	0	0
1	1	0	0	0
1	1	0	0	0
0	0	0	0	0
0	1	0	0	0
1	1	0	0	0
1	1	0	1	0

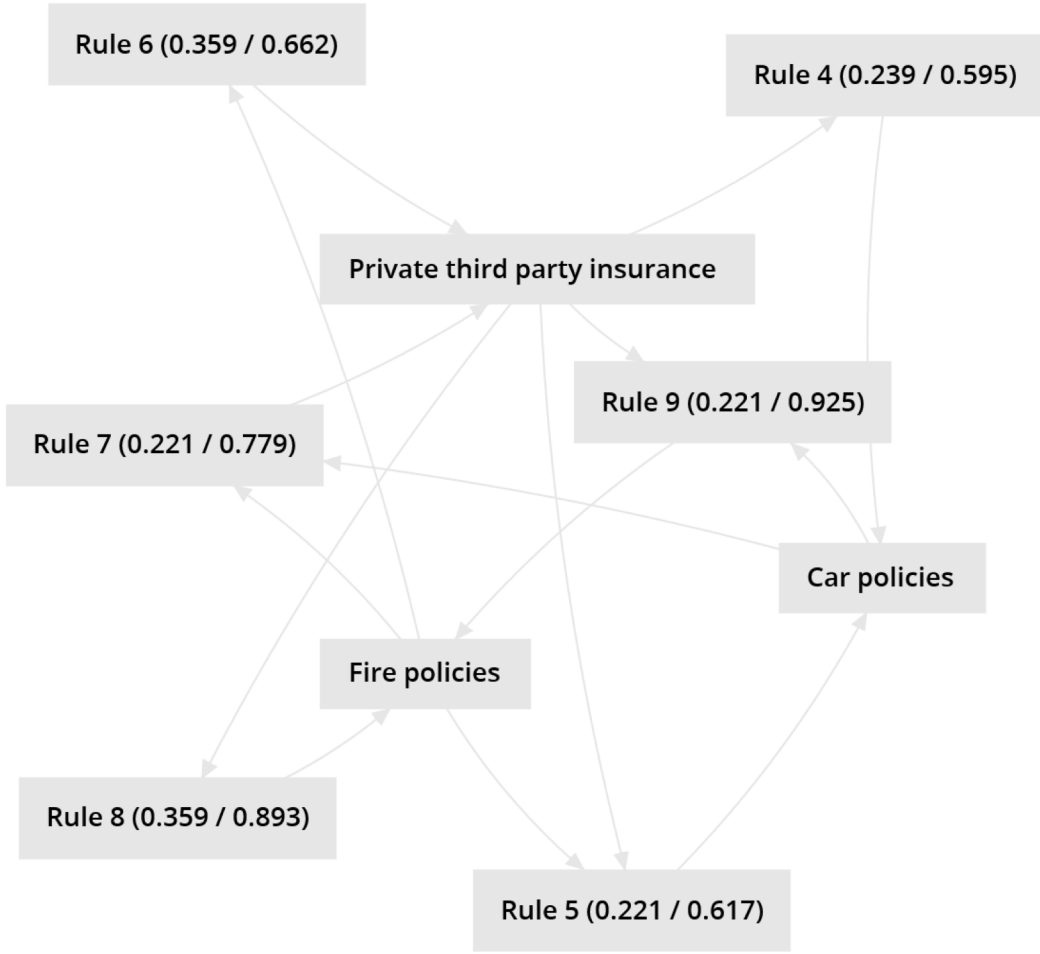
PROCESS IN RAPIDMINER

- Used the FP growth operator to count the frequency of each type of policy
- Association rules create rules based on the frequency of each policy



FREQUENCY & VISUAL REPRESENTATION OF THE RULES

Item 1	Item 2	Support	Frequency
Fire insurance	Car insurance	0.284	1653
Fire insurance	Private third party insurance	0.359	2090
Car insurance	Private third party insurance	0.239	1391



ARM output

We got an output with 9 association rules

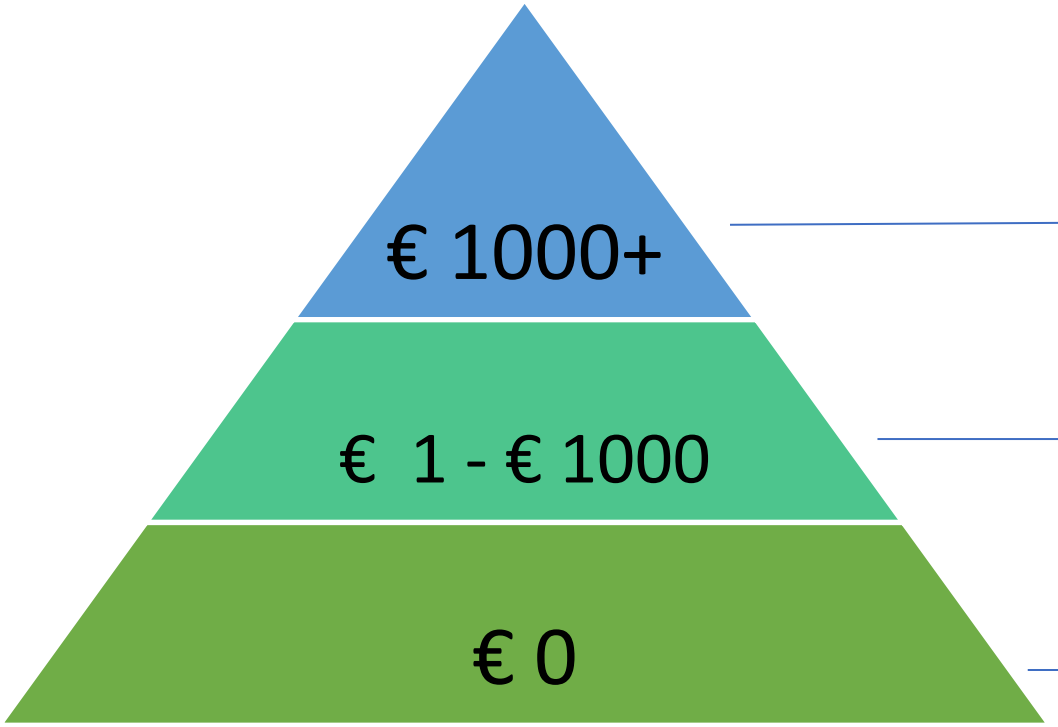
AssociationRules

```
Association Rules
[Fire policies] --> [Car policies] (confidence: 0.524)
[Private third party insurance] --> [Fire policies, Car policies] (confidence: 0.551)
[Car policies] --> [Fire policies] (confidence: 0.556)
[Private third party insurance] --> [Car policies] (confidence: 0.595)
[Fire policies, Private third party insurance] --> [Car policies] (confidence: 0.617)
[Fire policies] --> [Private third party insurance] (confidence: 0.662)
[Fire policies, Car policies] --> [Private third party insurance] (confidence: 0.779)
[Private third party insurance] --> [Fire policies] (confidence: 0.893)
[Car policies, Private third party insurance] --> [Fire policies] (confidence: 0.925)
```

ARM – INSIGHT IDENTIFIED

There exists a business opportunity to cross-sell the other two insurances the moment a customer buys Car insurance, Private third party insurance and Fire insurance

IDENTIFYING HIGH VALUE CUSTOMERS



Assumption

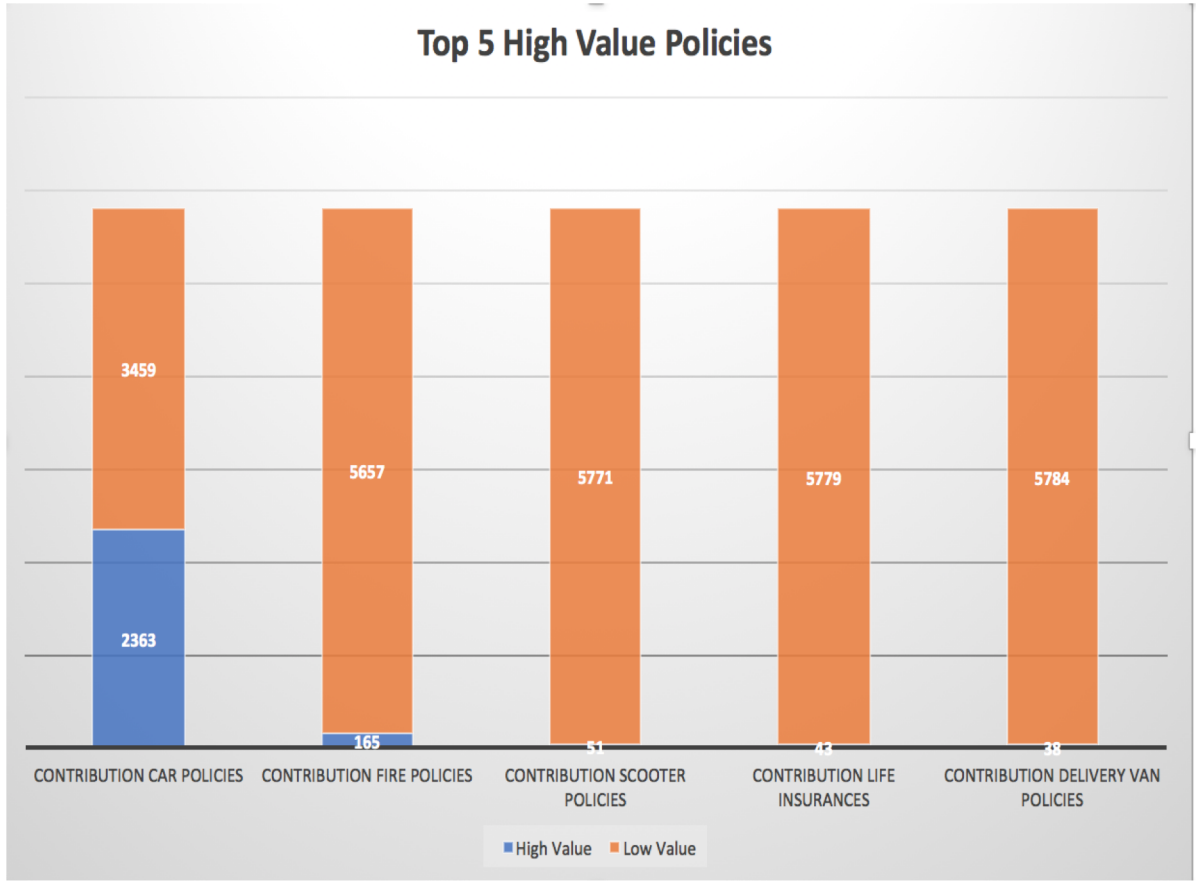
High Value Contribution

Low Value Contribution

No Contribution

Looked at contribution to policies variables (20 variables) and observed the top 5 high-contributors

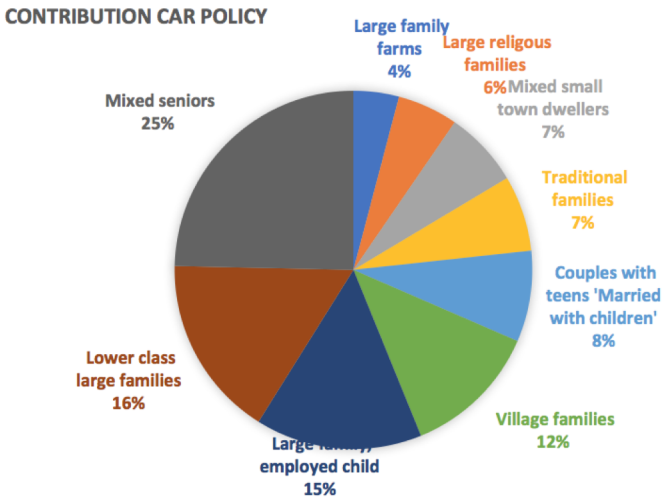
IDENTIFYING HIGH VALUE CUSTOMERS

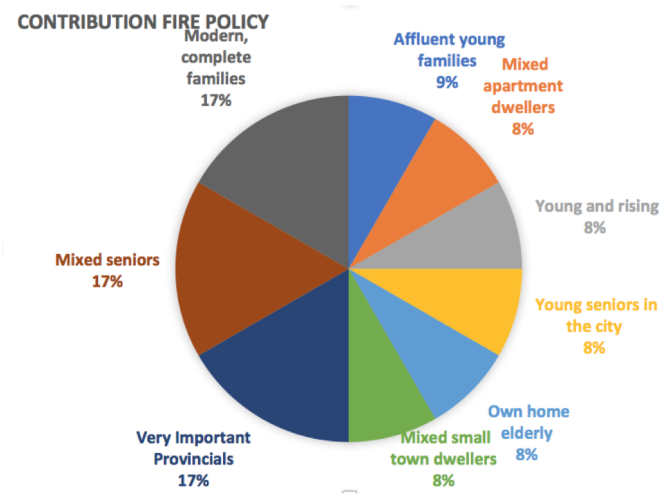


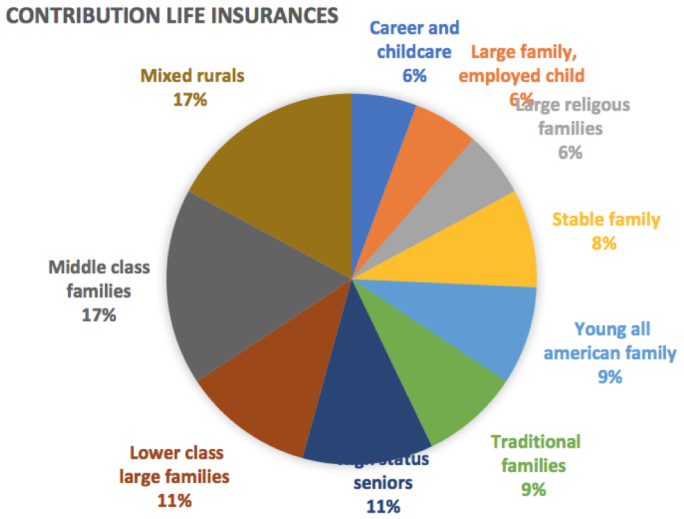
Top 5 Policies with € 1000+ Contribution

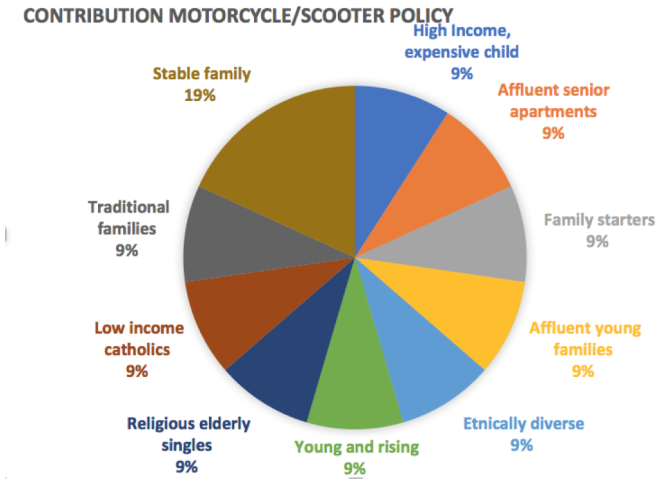
1. Contribution Car Policies
2. Contribution Fire Policies
3. Contribution Scooter Policies
4. Contribution Life Insurance Policies
5. Contribution Delivery Van Policies

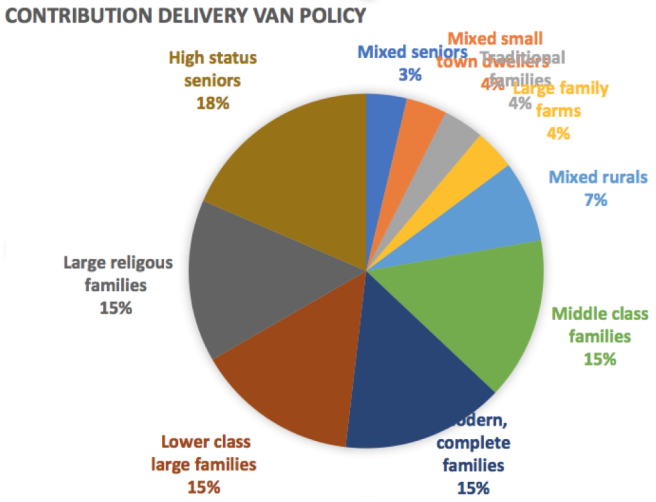
IDENTIFYING HIGH VALUE CUSTOMERS











IDENTIFYING HIGH VALUE CUSTOMERS

Results:

1. Mixed Seniors
2. Traditional Families
3. Large Religious Families
4. Mixed Rural
5. Career and Childcare

CONCLUSION

- Using Naive Bayes - accurately predict people buying caravan insurance policy using socio-demographic information of households

- Using association rule mining we have discovered cross selling opportunities for the insurance company

- Using data visualization, we identify high value customers for better customer maintenance


Thank
you!!