Collaborative Filtering for Life Insurance Products

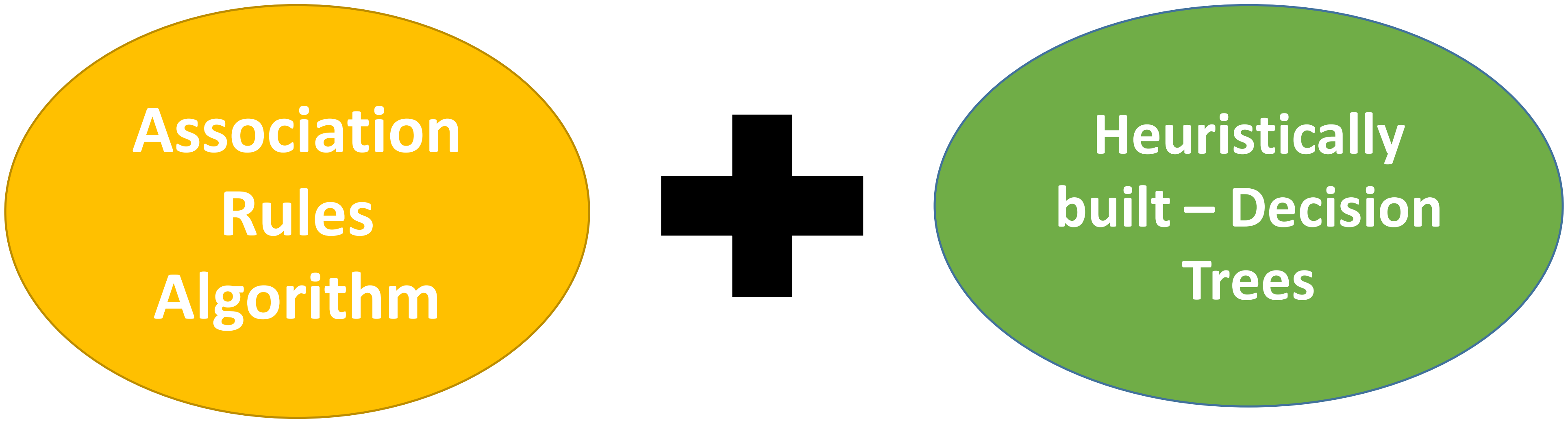
Jairo Fernando Gudiño-Rosero

Sample of Work

Philosophy

Usual collaborative-filtering algorithms (Non-Matrix Factorization, Weighted Alternating Least Squares, Clustering, SVD, Deep Learning) cannot address information as correctly as simple heuristics when information of the objective is too low.

I build a collaborative-filtering algorithm based on:



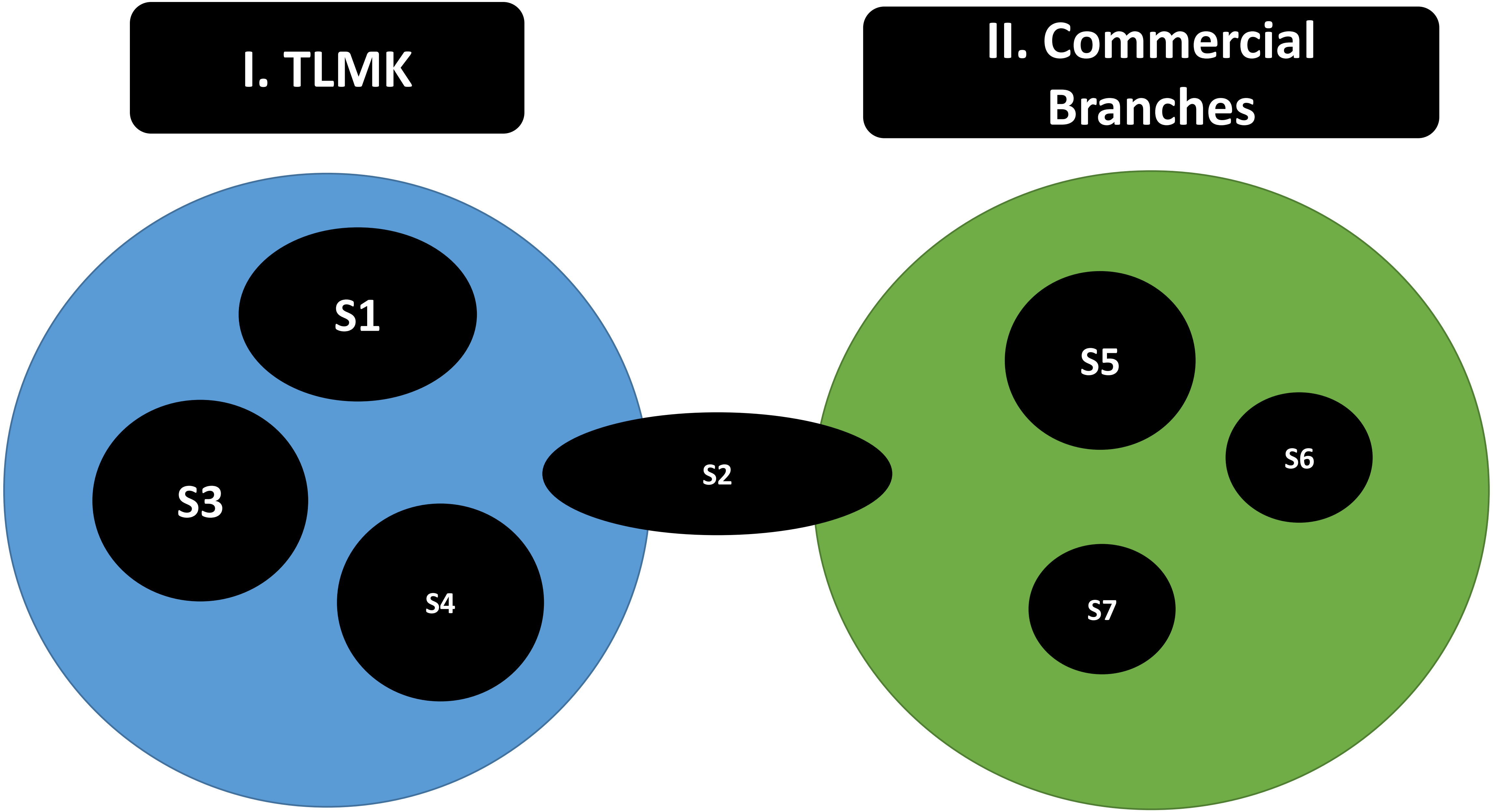
Business Context

What is the next best offer?

You are hired to build a collaborative-filtering algorithm for a life-insurance company in Latin America, but the market size is not too big and you must allocate 7 products with few available infomation.

Too limited information:

- **Channels:** TLMK (Call-center) and Commercial Branches. If the customer uses both, you can offer 2 products at most.
- **Features (almost not related):** Income, total wealth, age, economic sector, state, if is retired, skilled-degree of the job, type of job-contract, relationship status, if has a mortgage credit, if has an account in large banks & employment status.
- **Observations:** 88.000.


I. TLMK
II. Commercial Branches
S1
S3
S4
S2
S5
S6
S7

Statistical Facts

Statistical Facts: Association Rules

1. **TLMK**

- 81.6% (6.255) of educational sector-workers bought S1. 24.02% of total S1 costumers.
- 67.5% (3.648) of armed-forces listed people bought S2. 22.66% of total S2 customers.
- 47.2% (233) of retired people and living in the coastal area bought S3. 2.5% of S3 customers; the rest of retired people bought mostly S1 (80.1%, 3.646), being the 14% of total S1 customers.

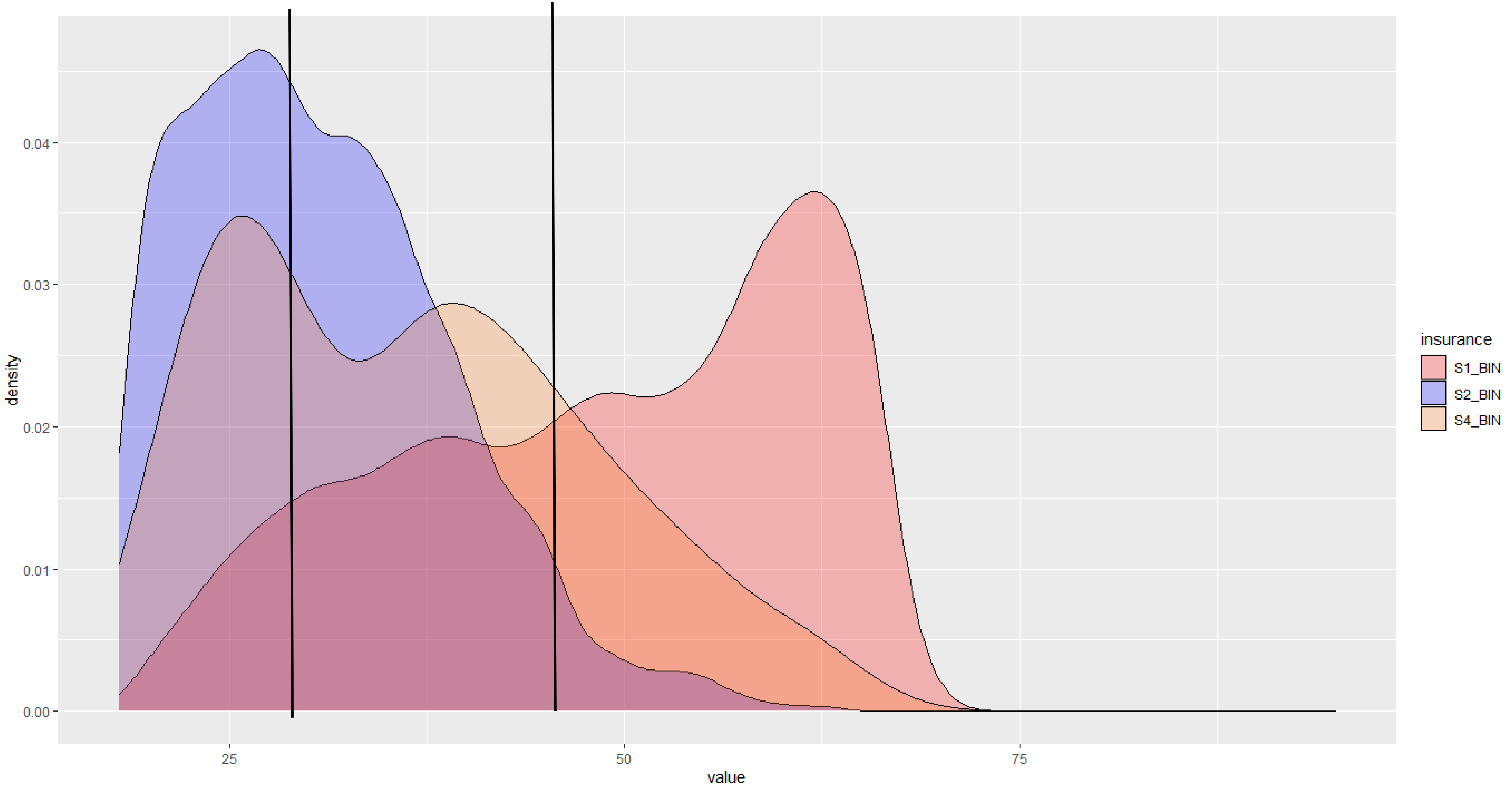
Statistical Facts: Association Rules

2. Commercial Branches:

- 66.7% (4.614) of retired people bought S5. 49.7% of total S5 customers.
- 71.2% (10.767) customers working in health, construction, social services, public administration or specialized professionals or having an active mortgage credit or receiving monthly wages in large bank accounts or with temporary-contract jobs or in a free union relationship bought S7. 35.61% of total S7 customers.
- 70.5% (508) of armed-forces listed people bought S2. 21.47% of total S2 customers.

Statistical Facts: Empirical Distributions

Age:



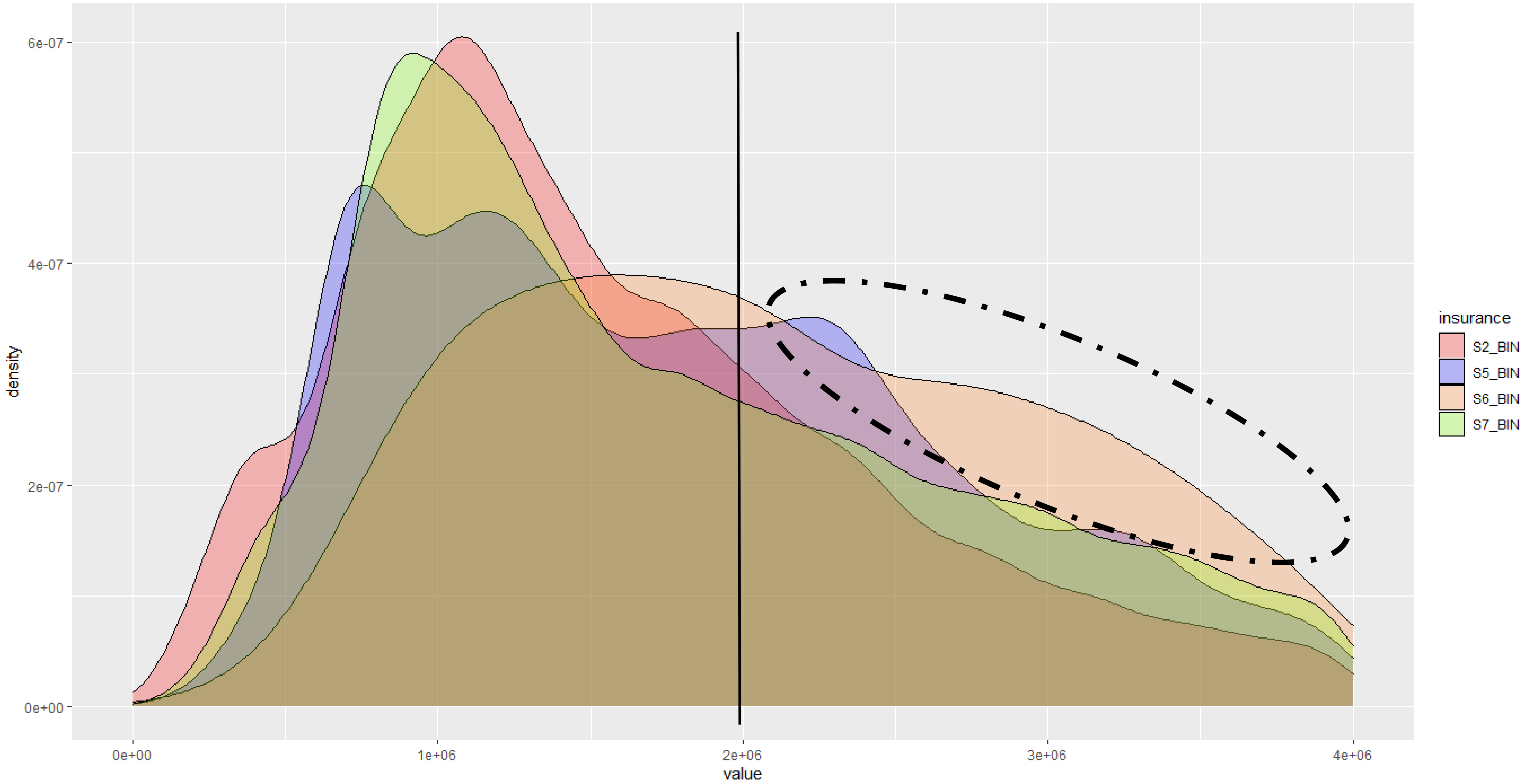
Statistical Facts: Empirical Distributions

In TLMK:

- After some filterings, decide the insurance to recommend according with the age of the customers: if age < 28, S2; if 28 <= age < 41, S4; if age >= 41, S1.

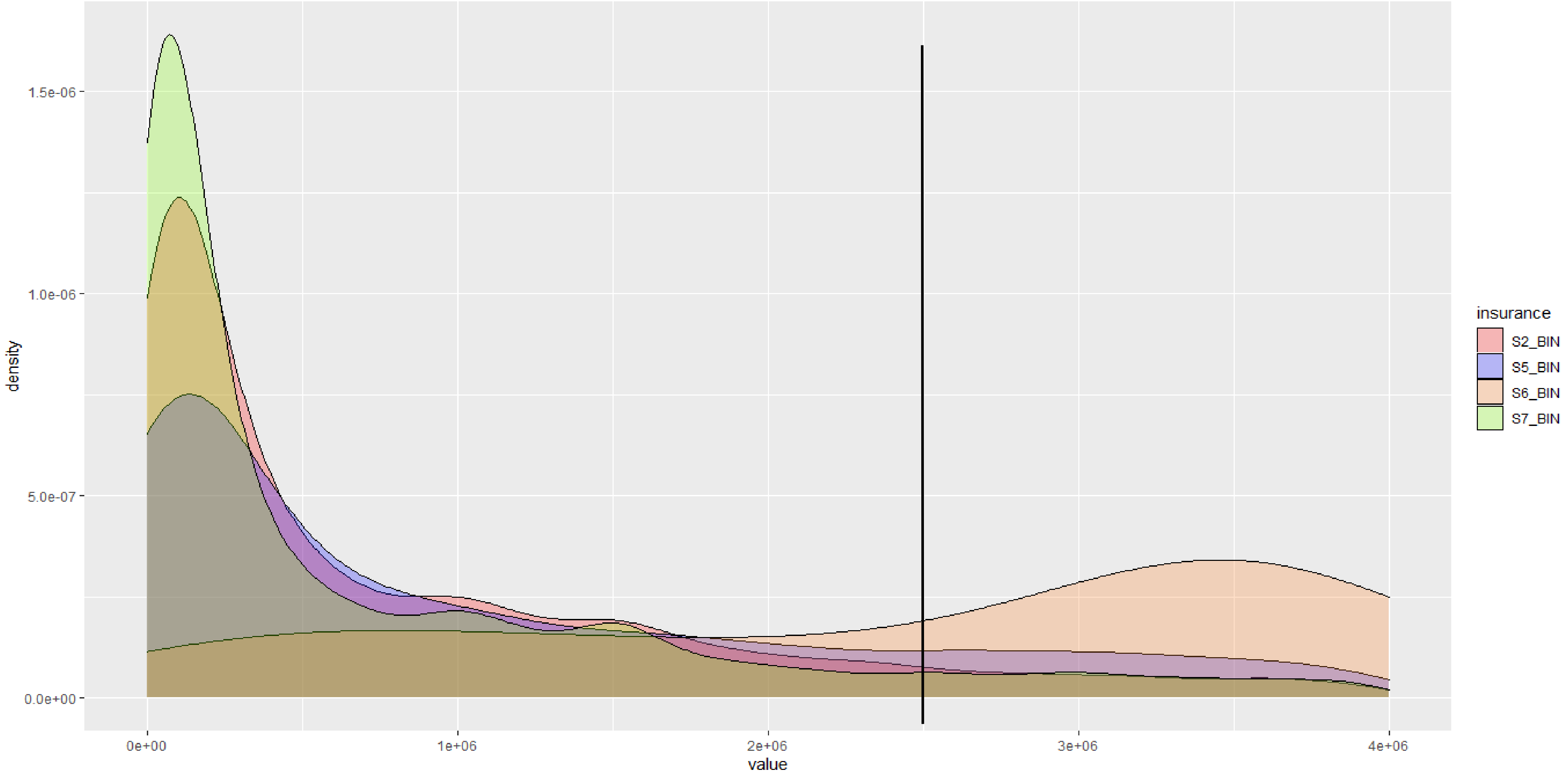
Statistical Facts: Empirical Distributions

Monthly Income:



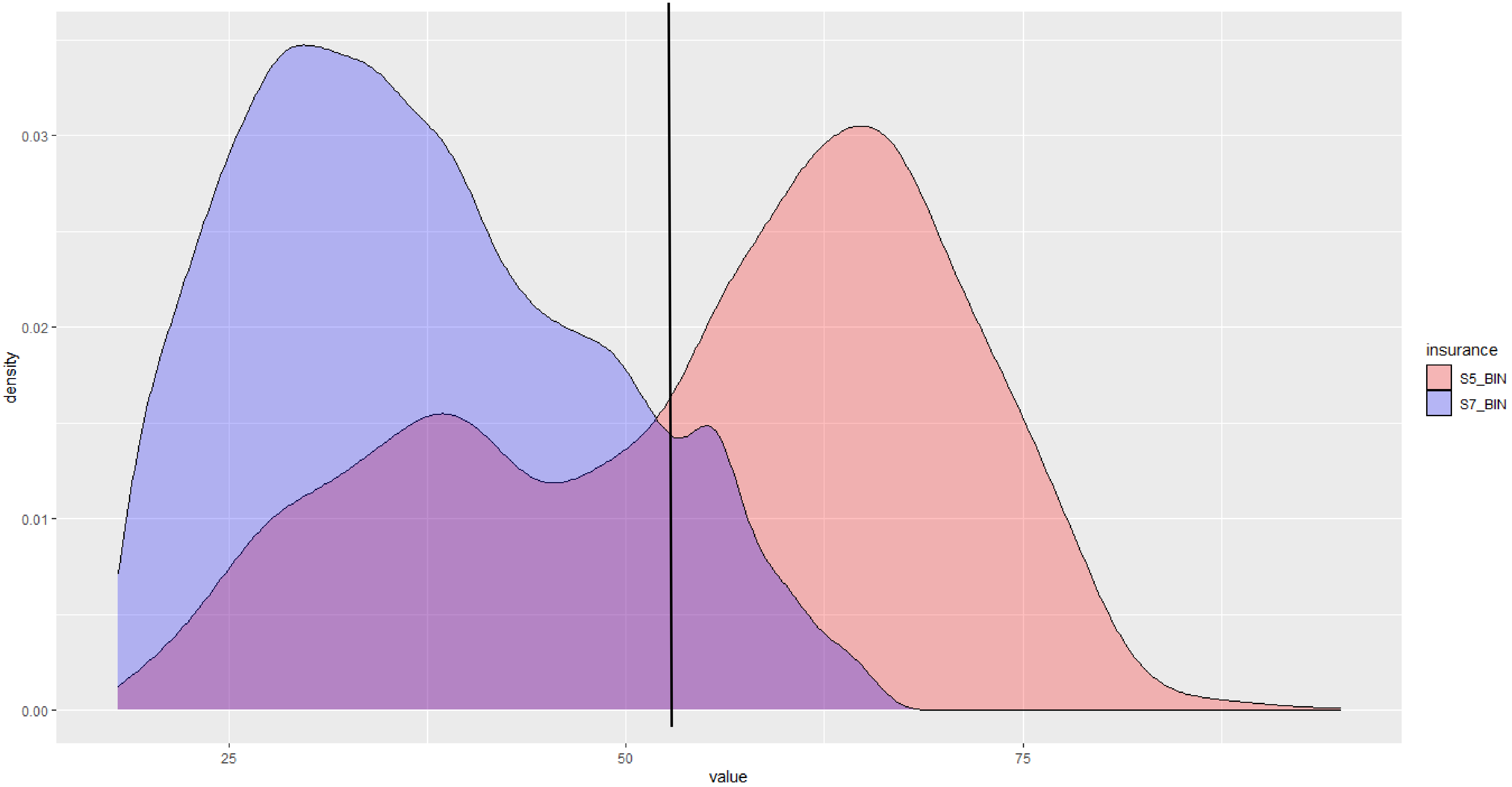
Statistical Facts: Empirical Distributions

Total Wealth:



Statistical Facts: Empirical Distributions

Age:



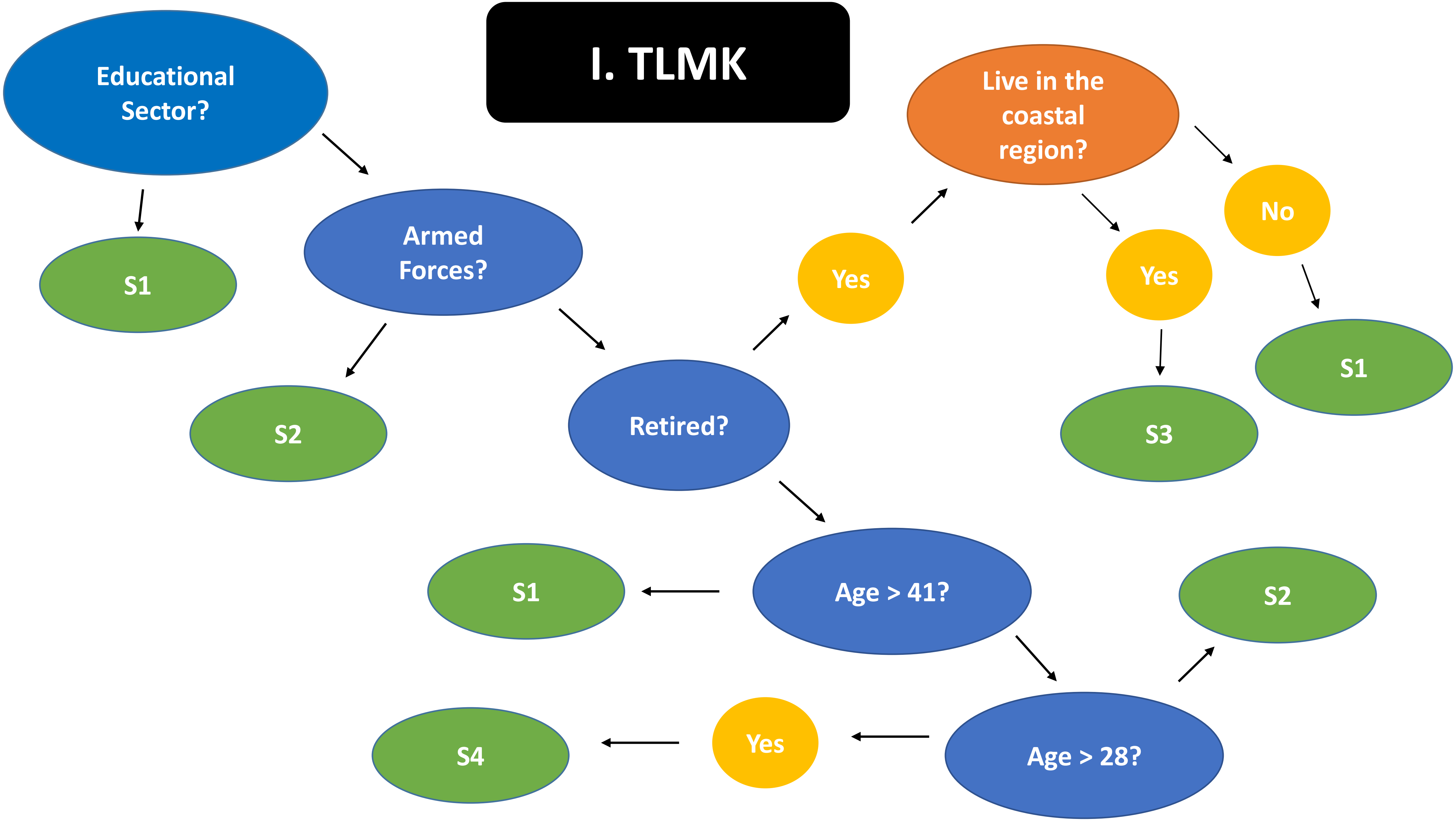
Statistical Facts: Empirical Distributions

In commercial branches:

- After some filterings, for customers with monthly income > 2M & total wealth > 2.5M, recommend S6 (2.5%, 170). This population is the 58.02% of S6 total customers.
- After some filterings, for customers with age > 55, S5 (53.7%, 476), encompassing 5.2% of S5 customers; otherwise, S7 (56.9%, 11.866 customers), encompassing 39.31% of S7 customers.

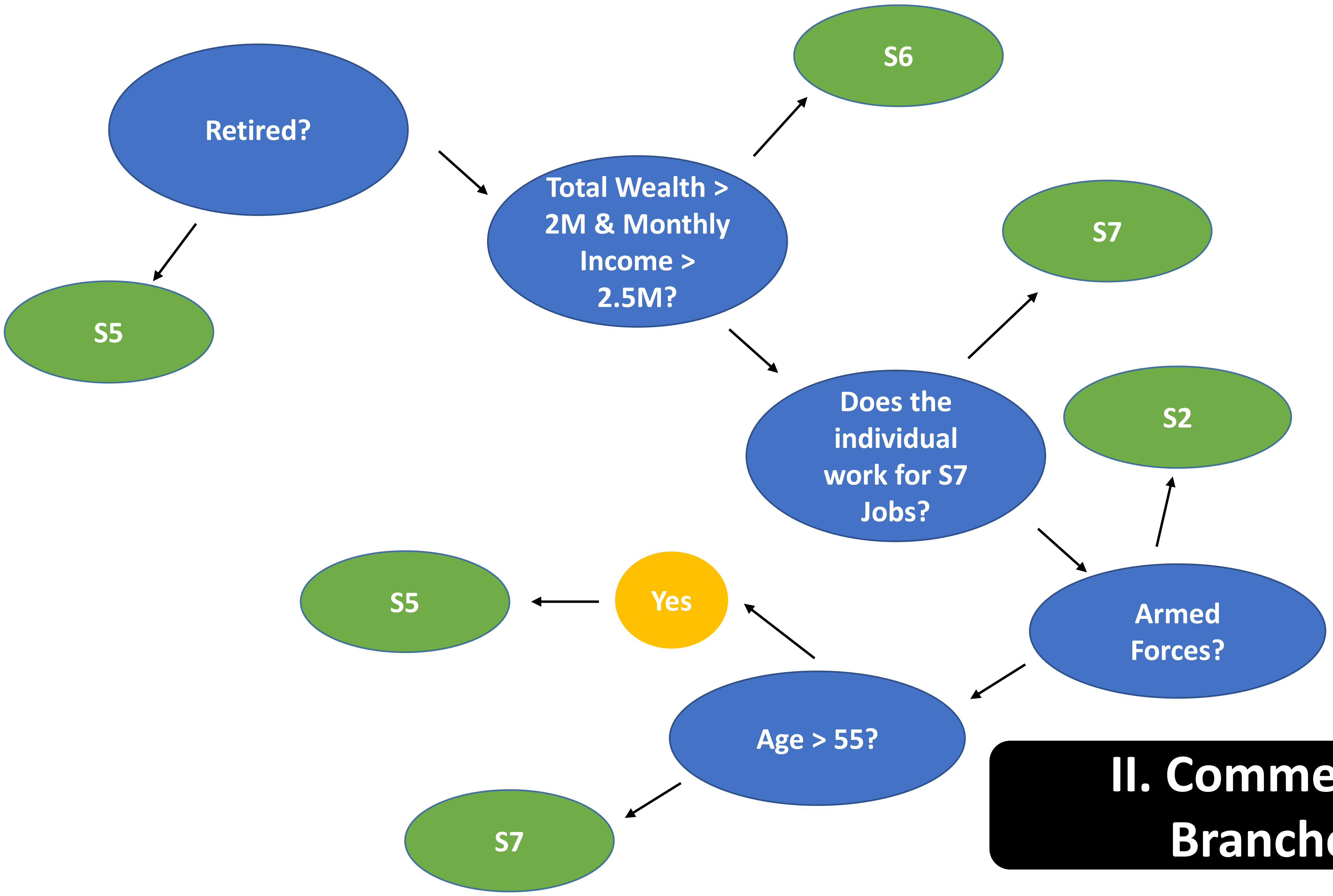
Results:
Heuristical Decision Tree

I. TLMK



Educational Sector?
S1
Armed Forces?
S2
Retired?
Yes
Live in the coastal region?
Yes
No
S3
S1
Age > 41?
S1
S2
Age > 28?
Yes
S4

II. Commercial Branches



Overall accuracy:

Low-medium overall accuracy, medium-high accuracy within clusters identified.

*TLMK:

- S1: 68.29%
- S2: 68.57%
- S3: 2.5%
- S4: 30.59%

* Commercial Branches:

- S2: 29.58%
- S5: 53.58%
- S6: 69.62%
- S7: 53.70%

Comparing other methods

I convert statistical facts to dummy and numerical variables and compare results with ML algorithms. I find is not straightforward for them to identify small differences, obtaining the accuracy power is not much better. Ensemble algorithms may be fruitful for calculating propensity scores and improve results.

Random Forests (with cross-validation and upsampling)

*TLMK:

- S1: 81.74%
- S2: 83.63%
- S3: 4.74%
- S4: 0%

* Commercial Branches:

- S2: 22.89%
- S5: 27.9%
- S6: 0%
- S7: 44.57%