Car Insurance

Jan Tomášek
Štěpán Havránek
Michal Pokorný
Competition details

Jan Tomášek
As a customer shops an insurance policy, he/she will receive a number of quotes with different coverage options before purchasing a plan.

Using a customer’s shopping history, can you predict what policy they will end up choosing?
Evaluation

Submissions are evaluated on an all-or-none accuracy basis. You must predict every coverage option correctly to receive credit for a given customer. Your score is the percent of customers for whom you predict the exact purchased policy.
Prizes

- First place: $25,000
- Second place: $15,000
- Third place: $10,000
Data structure

customer_ID, record_type, dateTime, location, group_size, homeowner, car_age, car_value, risk_factor, age_oldest, age_youngest, married_couple, C_previous, duration_previous, A,B,C,D,E,F,G, cost
## Product options

<table>
<thead>
<tr>
<th>Option name</th>
<th>Possible values</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0, 1, 2</td>
</tr>
<tr>
<td>B</td>
<td>0, 1</td>
</tr>
<tr>
<td>C</td>
<td>1, 2, 3, 4</td>
</tr>
<tr>
<td>D</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td>E</td>
<td>0, 1</td>
</tr>
<tr>
<td>F</td>
<td>0, 1, 2, 3</td>
</tr>
<tr>
<td>G</td>
<td>1, 2, 3, 4</td>
</tr>
</tbody>
</table>
Solution 0

- Last quoted plan benchmark
  - 53%
- don’t use exactly last quoted but average
  - weighted sum
  - deduce weights on train set
    - using genetic
      - based on user info
    - regression
      - based on column only
Our common interface

- Meta level script combining various solutions
  - BASH
- Aggregates solution’s outputs using their confidence flag
Weka

- Weka is a collection of machine learning algorithms for data mining tasks
- University of Waikato
- Very complex software
Weka live example

weather
Better features and problem reductions

Michal Pokorný
Better features, problem reductions

- Exponential distribution of plans
  - 283 plans w/ >100 purchases, 1700 plans total
- Most customers (~73%) choose some offered plan
- Some features seem less relevant
  - Time
Ideas to try

- Basic classifier trained on crude features was no better than naive solution
  - Naive: always pick the last plan
- Benchmark other naive solutions
  - Weight plan features by how many times were they picked, etc.
- Gain insights to meaning of individual plan properties
Ideas to try

- Train classifiers on mutilated original training data
- How many customers change their properties during the quoting process?
- Train a classifier just to decide when to use naive heuristic (performance ~53%)

Implementation: scikit-learn (Python)
Unsupervised learning approach

Štěpán Havránek
Unsupervised learning

- Mix of
  - Clustering
  - Evolution (genetic programming)
Clustering

- Somehow split the data items into categories
- Each category stands for one output
- New item is categorized and sets its output according to its category
Clustering
Clustering - our case

- Large input dimension
- Big value range of some input dimensions
- Not always ordered set
  - Enums
  - Date/Time
  - Geographic data
- Quite large output dimension
  - 7 output variables (ranges between 2 - 4)
Clustering - customization

- Choose only some features
- Overridden metric
  - Weighted distance for each dimension
  - Own ordering
    - Binary metric
    - Proprietary total order
Clustering - customization

● Output
  ○ Clustered categories for group of outputs instead of one particular output
    ■ Particular output will be decided by aggregation of category outputs
  ○ = Classification -> characteristic vector

● Output can carry information about its certainty
Clustering - categorization

- K nearest neighbours
  - Parameter K
  - Static/dynamic version

- M means (gravity centers)
  - M is given by number of categories we want to differentiate
  - Static/dynamic version

- Hierarchical clustering
Clustering - model complexity

- Our model is quite complex
  - A lot of parameters
    - Categorization technique
      - Its parameters
    - Feature weights
    - Own metrics
    - Output policy
- How to guess this parameters?
  - Tryout
  - Let the evolution do the work
Clustering and genetics

- Population member
  - Vector of numeric values
    - Weights
    - Parameters for categorization technique
  - Enum values
    - Categorization technique
    - Output aggregation type
Q & A