Enova Data Smackdown 2013

Nelson Auner, Zi Chong Kao

Analysis of Problem

Original Task:

Select customers to contact to maximize loan profit less calling cost

Analysis:

- Calling costs constant at 10 Euros
- Other costs to acquire customer assumed to be negligible
- Max loan profit achieved by contacting all customers with positive expected net value

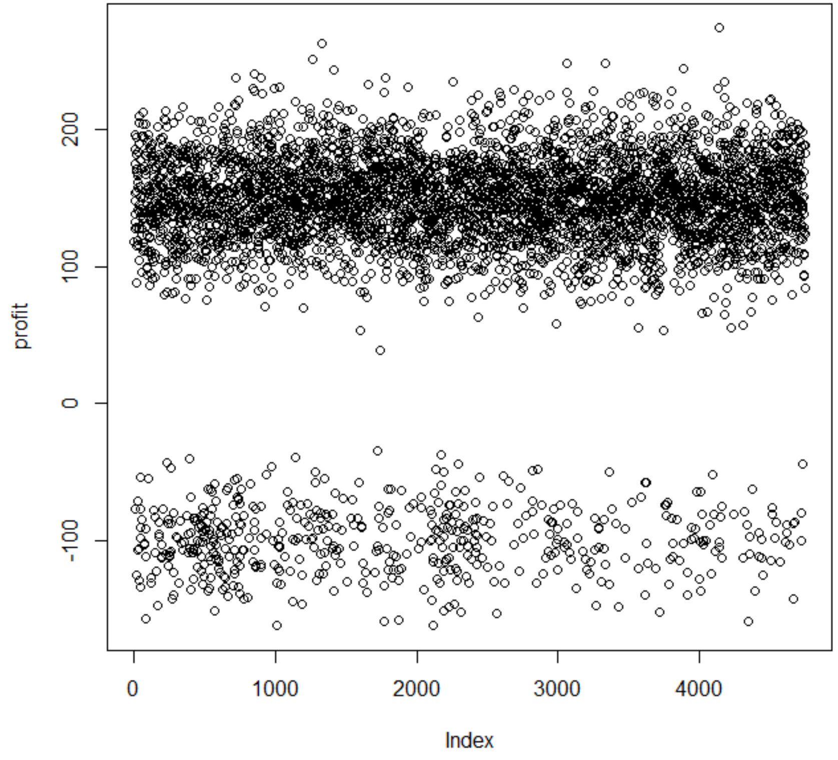
Reframed Task:

Determine customers with expected loan profit > 10 Euro

Data Understanding

Loan profitability is binary rather than continuous

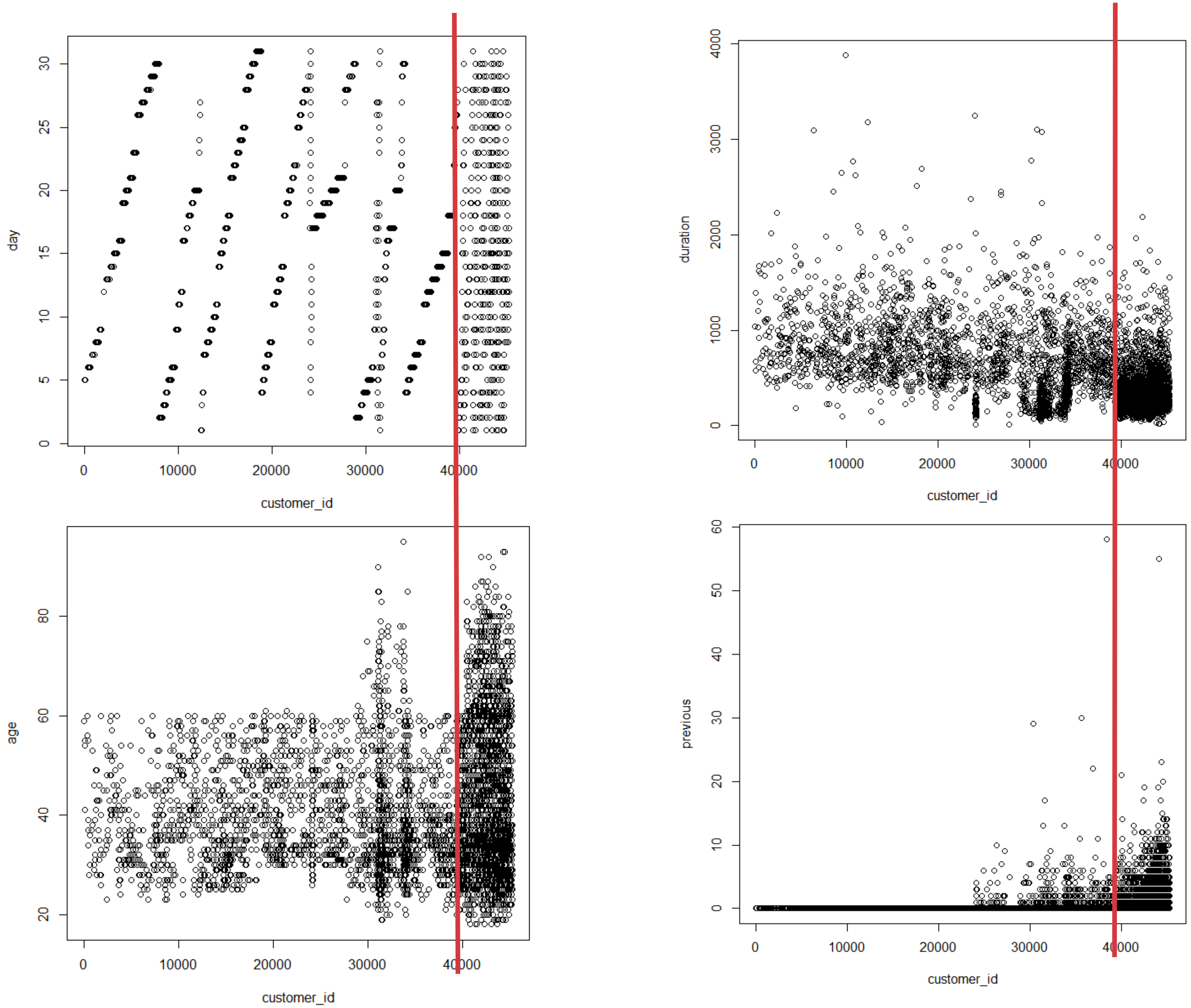
Observe distribution on the left. Given that threshold for probability is 10 Euro, customers are either clearly profitable or not.



Data Understanding

"Historical" dataset is actually a mixture of 2 different datasets!

We speculate that customer_ids <40 000 and >40 000 are from different datasets. Observe the different distributions on multiple covariates:



Model Overview

<u>Goal</u>

Determine E(loan profit) for each callee

<u>Process</u>

Given that expected profit is only obtained in two stages:

Response → Loan Profitability → Expected Profit

We decompose E(loan profit) in the following way:

$$E(\text{loan profit}) = E(\text{loan profit} \mid \text{respond} = 1)\ P(\text{respond} = 1)$$

<u>Analysis</u>

- Decomposition important to study responders and non-responders separately.

- Otherwise, incorrectly model loan profit for non-responders as 0: Non-responders could be profitable if they had taken out a loan

2 Stage Model

Recall: E(loan profit) = E(loan profit | respond = 1) P(respond = 1)

Response Model:
- Determine P(respond = 1)
- Output is a probability rather than indicator variable:
Allows customers with low response probability to be selected if they have high enough expected loan profitability
- Logistics regression

Profitability Model:
- Determine E(loan profit | respond = 1)
- Given our earlier observation that loan profit is binary, we use a logistics regression model as well

Data Preparation

Response Model

- Clean data removing incorrect cases (eg. Age > 100)
- Recoding categorical variables: Easier to work with 1,0 than “yes”, “no”
- Using the 10% of our data set aside as a test dataset, we wrote a script to run our strategy on the data.

Profitability Model

- Ensure no data is obviously wrong
- Separate out the 4748 of 40712 customers who responded

Evaluation of Results

Response Model

- Root Mean Square Error = 0.249

Profitability Model

- Error = 0. Perfect performance!
- Our logistics model predicted all 4748 of the responder's profitability correctly.
- Consequently, Response Model now bears all the load.

Overall

- Profit of our strategy was more double (2.37x) the profit of the naïve strategy of targeting every customer.
- Our model's profit: 47 308.30
  Naïve model profit: 19 992.55

Recommendations and Future Research

Separate the two datasets out more finely.
Here, we visually inspected that the cut off was at customer_id 40 000. It's probably slightly less than that. We also note that customer_ids near 25 000 and 32 000 display very similar behavior to the >40 000 group. Ideally, we'd group all of these data together and conduct separate analysis.

Examine how the two datasets are different from each other.
We suspect that they were derived using very different methodologies. The customer_id <40 000 seem a good deal more systematic and "clean". The days show clear sequential structure, and the ages are cut off between 22 and 60.

Use classification algorithms (eg. K-nearest neighbors).
Given that loan profitability is binary with distinct clustering, we should treat this as a classification problem rather than a regression problem. Classification is after all, much easier than regression