Undercutting Based on Analytics

Background

In most industries, costs of inputs can be used to calculate a profitable price to charge for a product. In the insurance industry, their product is the “insurance policy”, and the largest input is the cost of “claims” their customers file when they get into an accident. These costs are difficult to calculate for two reasons. First, they vary by customer; some customers have multiple claims over the life of their policy, others have none. Second, the total cost of the claims is not known until after an accident occurs, which is well after the policy has been priced and sold to the customer.

However, segments of customers may be more likely to cause accidents the other (e.g. young males are much riskier than middle-aged females) leading insurance companies to charge different prices to various groups to “out-segment” their competitors leading to adverse selection for the competitors who end up with a large number of high-risk and claim-prone customers.

Insurance companies tightly guard the algorithms that they use to price the policies of individual customers, but if we could provide a policy quote after observing our competitors quote, this would allow for an “instantaneous last mover” advantage.

Several sites provide services of comparing quote and considerable variability is observed between competitors. At that point it is already too late to update your quote. (Additional legal issues further complicate trying to craw competitors’ websites in real time.) Therefore, if we can predict competitors’ (dynamic) quotes we can base our quote on those as well.

You were approached by an auto insurance company, which was interested in pursuing such idea based on analytics. They ask you to conduct a pilot study, comment on the feasibility of the project, and potential next steps.

Moving Forward

As the starting point, a data file containing the information of 15,484 quotes of an online competitor (ALLSTATE) was provided. The file contains information about the policy offered, information on the customer, and the quoted price (see Exhibit 1 below for a description). As a last reminder to potentially help your modelling, the manager discussed with you how prices are typically quoted in the insurance industry. In particular he stressed that many features appear as add-on prices.

Exhibit 1 - Variable Descriptions

customer_ID - A unique identifier for the customer
shopping_pt - Unique identifier for the shopping point of a given customer
record_type - 0=shopping point, 1=purchase point
day - Day of the week (0-6, 0=Monday)
time - Time of day (HH:MM)
state - State where shopping point occurred
location - Location ID where shopping point occurred
group_size - How many people will be covered under the policy (1, 2, 3 or 4)
homeowner - Whether the customer owns a home or not (0=no, 1=yes)
car_age - Age of the customer's car
car_value - How valuable was the customer's car when new
risk_factor - An ordinal assessment of how risky the customer is (1, 2, 3, 4 or NA)
age_oldest - Age of the oldest person in customer's group
age_youngest - Age of the youngest person in customer's group
married_couple - Does the customer group contain a married couple (0=no, 1=yes)
C_previous - What the customer formerly had or currently has for product option C (0=nothing, 1, 2, 3,4 or NA)
duration_previous - how long (in years) the customer was covered by their previous issuer (duration or NA)
A,B,C,D,E,F,G - the coverage options:
A = Collision (levels: 0, 1, 2);
B = Towing (levels: 0, 1);
C = Bodily Injury (BI, levels: 1, 2, 3, 4);
D = Property Damage (PD, levels 1, 2, 3);
E = Rental Reimbursement (RR, levels: 0, 1);
F = Comprehensive (Comp, levels: 0, 1, 2, 3);
G = Medical/Personal Injury Protection (Med/PIP, levels: 1, 2, 3, 4)
cost - cost of the quoted coverage options (i.e., ALLSTATE's quote to the customer)

Case 1: Undercutting Based on Analytics

Instructions

This is a team assignment. Each member of the team receives the same grade. Submission is online (see course webpage). In order to be graded, you need to upload one pdf file (no longer than 4 pages with font size 12pt) and your R script (this should be well commented and run without errors). Any additional material you judge relevant that complements your submission can be submitted as additional files. Make sure that the section number and all names of the team members are clearly listed. Late submissions (but submitted before in-class discussions) or inappropriately formatted cases will have points deducted. Missed cases are worth 0 points. <u>Important: submit a PDF file and follow the naming convention (see Syllabus).</u>

Assignment

Your answers should be clear and provide unambiguous recommendations when asked. Please provide explanations for your answers and any outputs that you feel are needed to support your argument. Keep in mind that this is an open ended exercise. There is no exact model of reality as discussed in class. You will be evaluated on your modeling of the problem, judgment of which core task to use, appropriateness of the choice of algorithm, and taking all of that to data. (Therefore do not stress about trying to fine tune details.)

Assumptions are needed to implement any estimation strategy. Initially, assume that ALLSTATE does not differentiate customers through customer ID, number of visits, time of the day, and location (ZIP code).

Questions

1. Pick two (or more) variables and attempt to show a relation between them via visualization. As discussed before, this requires one to formulate a question, and to communicate clearly a conclusion based on data visualization (specify the why, what, how).

2. Provide a model based on linear regression to forecast the quoting procedure from ALLSTATE based on the observed variables. Pick two variables of your model, describe their marginal impact on the quote, and comment the interpretation from the business perspective.

3. Suppose that a customer will pick the lowest between the quote you provide and that ALLSTATE provides. Build a model framework (follow/adapt steps in Model Framework in Class 3 for the Churn Problem) to *maximize expected revenue*[1] from a customer given the observed characteristics. This includes the mathematical model, description of a decomposition strategy, the associated core tasks, and specific data mining methods you would choose. For each core task comment if it can and if it cannot be implemented with the available data.

4. Suppose that a customer will pick the lowest between the quote you provide and that ALLSTATE provides. Aiming to *maximize expected revenue*[2], provide quotes for each of the three customers specified in “new.customers”. Clearly state which core task and which data mining method you used to provide the quote.

5. Suppose next that the customer might not accept either of the two quotes (but he will consider only the smallest of the quotes). Build a model framework (follow/adapt steps in Model Framework in Class 3 for the Churn Problem) to *maximize expected profit* from a customer given the observed characteristics. This includes the mathematical model, description of a decomposition strategy, the associated core tasks, and specific data mining methods you would choose. For each core task comment if it can and if it cannot be implemented with the available data.

[1] Immediate revenue, not lifetime revenue.

[2] Immediate revenue, not lifetime revenue.