Kellogg School of Management
Northwestern University

Pentathlon III:
Next Product to Buy Modeling

The e-mail frequency test changed the dynamics of the monthly Product Department Director meetings.[1] The test put to rest the question of whether limiting the number of promotional e-mails to customers was in the best interest of the company. The test had also increased Anna Quintero's credibility in digital marketing. Most surprising to Anna, however, was that the department directors had started to seek her advice on problems that required customer analytics.

Such a problem had been at the center of a recent meeting. In fact, it was a problem of Anna's own making. The company-wide agreement to limit promotional e-mails to two per week meant that the different departments now had to coordinate their promotional e-mail activities. Having to coordinate was not in and of itself something the department directors objected to. They were used to negotiating over scarce resources such as marketing budgets, retail space, and head counts. Instead, the problem was a lack of information. François Cabret, the department head for Endurance Sports, put it this way:

> "When we negotiate over budgets or space allocation in our stores there are some important metrics we use. For example, we all agree that sales per square foot of store space is important. However, when we negotiate over how to allocate promotional e-mails across departments, we just can't see eye-to-eye. For example, I have argued that endurance-themed e-mails should have highest priority for women because running is just as popular among women as it is among men. But Patricia, the department head for Racquet Sports, says that, online, her category has particular sales appeal for women, even when this is not the case in stores. Frankly, I don't know how to resolve these questions. We can get reports about which customer segments buy which products online but that is **not** what we need. What we really want to know is how **effective** different promotional e-mail messages are for different customer segments. However, there are so many different customer types with different purchase histories that I don't even know how to start thinking about finding the answer."

During the meeting the department heads had discussed the idea of running another test in order to determine the effectiveness of different promotional e-mail messages for different customer

[1] Each department director runs one of the seven major product categories sold by Pentathlon: Endurance (e.g., running, cycling), Strength and Fitness (e.g., gymnastics, yoga), Water Sports (e.g., sailing, kayaking), Team Sports (e.g., soccer, basketball, rugby), Backcountry (e.g., hiking, climbing), Winter Sports (e.g., skiing, snowboarding), and Racquet (e.g., tennis, badminton).

Professors Florian Zettelmeyer and Song Yao prepared this case to provide material for class discussion rather than to illustrate either effective or ineffective handling of a business situation. The names and the data used in this case have been disguised to assure confidentiality.

segments. But the idea had fizzled because no one felt like waiting another eight weeks, the length of the previous e-mail frequency test, to get an answer. Ten minutes after the idea of a test had been rejected Anna suddenly spoke up:

> “We may have been thinking about this the wrong way. I agree, we should not run a test – but not because it is going to take too long. We don’t need to run a test because we already have all the data we need to figure out which promotional e-mail message works best for different customer segments. In fact, I think my analytics team will be able to do better than that. If I am not mistaken, we should be able to analyze the effectiveness of different promotional e-mail messages for individual customers, not just for broad segments. Give me a few days and I will get you some answers.”

The Idea

In the months since the decision to limit customer e-mails, the departments – unable to agree on an optimal allocation procedure – had used random allocation as an interim compromise. Anna’s sudden realization during the meeting was that the random allocation had created something close to experimental data that could be used to analyze the effect of different promotional messages.

The allocation had been implemented as follows:

- Each week the digital marketing department split customers with valid e-mail addresses into seven randomly assigned e-mail groups.
- Each of the seven departments was allocated one of these e-mail groups for their exclusive use during that week, subject to the e-mail frequency limit.
- The e-mails sent by each department would be designed by that department and would feature products from that department. Of course, once customers clicked on the promotional e-mail and were on the Pentathlon website they could buy products from any department or none at all.

While this procedure had been chosen because it did not favor one department over another and because it was easy to administer, Anna noticed that it was ideally suited to analyze how different customers reacted to different messages. The key was that customers were being allocated to departments – and therefore to differently themed messages – randomly.

The Data

Anna asked her analytics team to pull the following data:

- The data pull should be based on the last e-mail sent to each customer. Hence, an observation would be a “customer–promotional e-mail” pair.
- The data should contain the basic demographic information available to Pentathlon:
    - “age”: Customer age (coded in 4 buckets: “< 30”, “30 to 44”, “45 to 59”, and “>= 60”)
    - “gender”: Gender coded as F or M
    - “income”: Income in Euros, rounded to the nearest 5,000 **€**
    - “education”: Percentage of college graduates in the customer’s neighborhood, coded from 0-100
    - “children:” Average number of children in the customer’s neighborhood

- The data should contain basic historical information about customer purchases, specifically, a department-specific frequency measure.
  - "freq_endurance – freq_racquet": Number of purchases in each department in the last year, excluding any purchase in response to the last email.
- The key outcome variables should be:
  - "buyer": Did the customer click on the e-mail and complete a purchase within two days of receiving the e-mail ("yes" or "no")?
  - "total_os": Total order size (in Euros) conditional on the customer having purchased (buyer == "yes"). This measured spending across all departments, not just the department that sent the message.
- *While of no importance for the prediction model,* Anna wanted to see from which departments the customer ordered when they purchased. This was captured in:
  - "endurance_os – racquet_os": Department-specific order size (in Euros). This was a breakdown of the total order size if buyer == "yes". The value was zero for most departments because customers rarely bought products from multiple departments on a single purchase occasion.

Finally, Anna requested that her team divide the data into a training sample and a test sample using a 70 – 30 split. She suggested 70,000 observations in the training sample and 30,000 observations in the test sample.

Anna requested the team sample more customers who purchased relative to those that did not purchase. She had learned that some analytical tools did a better job of scoring customers if the response variable had a similar number of "yes" and "no" values.

To achieve a 50/50 split between buyers and non-buyers for the training and test sample, her team randomly picked 50,000 buyers and added 50,000 randomly sampled non-buyers. The 100,000 customers were then randomly split into a training sample (70,000 customers) and a test sample (30,000 customers).

In addition to the 100,000 customers used for training and test, Anna asked the team to add a *representative sample* consisting of another 100,000 customers. This sample was representative in that it was a true random sample of the population and therefore contained the average proportion of buyers, namely 1%. This sample would be used to determine the expected benefits from using a next-product-to-buy model.

In summary, the dataset (see *pentathlon_nptb.pkl* in the git repo) contained 200,000 customers.

1. 70,000 in a training sample (**`training == 1`**)
2. 30,000 in a test sample (**`training == 0`**)
3. 100,000 in a representative sample
   (use **`is.na(training)`** or **`representative == 1`**)

Anna reminded the team that the predicted purchase probabilities from running a model on a dataset with 50% buyers and 50% non-buyers would be **much** higher than the actual response rate. This was because (most) models automatically ensure that the average of the predicted purchase probabilities corresponds to the actual response rate in the data used in estimation.

To scale the predicted purchase probabilities for use in the representative sample, the team could use (case) weights. The formula to generate the (case) weights is shown below:

```
pentathlon_nptb["cweight"] = rsm.ifelse(pentathlon_nptb.buyer == "yes", 1, 99)
```

See the documentation for statsmodels – glm for more information on how to estimate a model with (case) weights.

https://www.statsmodels.org/devel/generated/statsmodels.genmod.generalized_linear_model.GLM.html

The resulting prediction would have the correct magnitude for the representative sample and could be used for profit calculations.

Not all models have an option to specify case weights or may even perform poorly when weights are applied during estimation. For these models, the team could use pentathlon_nptb_5M.pkl. This dataset had 5M observations available for training and test and an extra 100,000 for the final predictions.

The downside to using the dataset with 5M rows is, of course, that estimation time will increase substantially. I do NOT recommend you use this dataset to select your final model or for tuning hyper parameters. You can, however, use this larger dataset to re-estimate your chosen model and generate profit estimates for the representative sample. You can download the 5M dataset through the link below.

https://www.dropbox.com/s/0i690ep4xcugi8t/pentathlon_nptb_5M.pkl?dl=1

Finally, Anna pointed out that reweighting should have no effect on the relative ranking of purchase probabilities across customers and/or offers and would only be relevant to determine the final projected profits.

The Analysis

After compiling the data, the digital marketing analytics team began to work through Anna's instructions:

Perform all estimation using the training sample. Use the test sample to assess model performance and check for overfitting. Perform the following calculations for each customer in the representative sample:

1. **For each customer** determine the message (i.e., endurance, strength, water, team, backcountry, winter, or racquet) predicted to lead to the highest **probability of purchase**. Describe your approach.
2. For each message, report the percentage of customers for whom that message maximizes their **probability of purchase**.
3. For each customer, determine the message (i.e., endurance, strength, water, team, backcountry, winter, or racquet) predicted to lead to the highest **expected profit** (COGS is 60%). Describe your approach to predict order size and how you calculated expected profit.
4. Report for each message, i.e., endurance, racket, etc., the percentage of customers for whom that message maximizes their **expected profit**.
5. What expected profit can we obtain, on average, per e-mailed customer if we customize the message to each customer?

6. What is the expected profit per e-mailed customer if every customer receives the same message? Answer this question for each of the seven possible messages (i.e., endurance, strength, water, team, backcountry, winter, or racquet).
7. What is the expected profit per e-mailed customer if every customer is assigned randomly to one of the seven messages?
8. For the typical promotional e-mail blast to 5,000,000 customers, what improvement (in percent and in total Euros) could Pentathlon achieve by customizing the message to each customer rather than assigning customers a message randomly?

A New Policy Proposal

In addition to presenting the results of the analysis during the next monthly department director meeting, Anna Quintero decided to propose a new process for allocating promotional e-mails across departments that was based on her team's analytical results. She wrote a draft for a new e-mail policy:

1. Promotional e-mails will be allocated to departments on a monthly basis
2. During a month e-mails will be assigned to departments as follows:
    a. For each customer, the analytics team determines the two messages that yield the highest expected profits
    b. The two departments whose messages yield the highest expected profit for a customer each control ½ of the allowed e-mail messages to that customer during that month
3. During the last week of each month the analytics team uses the data from e-mails sent during the first three weeks of that month and repeats the analysis described in step 2

Case Questions

1. Perform the analysis following the instruction e-mailed by Anna to the analytics team (Step 1 – 8 above). (**22 points**)
2. Comment on the new e-mail policy proposal. What are its weaknesses? Suggest at least one improvement? (**8 points**)
3. In addition to your Jupyter notebook with analysis and text, with your team, create a video presentation that lasts no more than 10 minutes and that covers your approach to building and finding the best model for this problem. Target your presentation to (1) the senior data scientist at the company and (2) the product manager for intuit QuickBooks. You should explain the technical details of your work and provide context on the proposed next steps in decision making. Please use Pentathlon III: Next Product to Buy Modeling (Video submission) to upload the link to your group video on Panopto (**10 points**)

Hints

- The variables endurance_os – racquet_os give a sense of how customers allocate spending across categories. However, this information should **not** be used in the predictive models.
- **Time-saver**: Your first priority in this case should be to generate (customized) predictions for each customer (step 1-8). Once you have achieved that objective, you can attempt to fine-tune your model(s) and try alternatives tools (e.g., a neural network or a tree-based

model). You can use the server (https://rsm-compute-01.ucsd.edu) to run cross validations. I recommend you use git to clone your code from GitLab onto the server.

- In this case you should select your final model, first and foremost, based on how well it fits the data. The more accurate the model, the better it will be able to predict the profit that can be achieved by customizing which message to send to which customer. Use model performance measures such as AUC, R-squared, RMSE, etc. to compare models.
- In Python, please adapt the code below for the full model you want to estimate. Note that you may not be able to even estimate this same model on the 5M row dataset due to inefficiencies in the estimation algorithm used in python.

```
import statsmodels.formula.api as smf
from statsmodels.genmod.families import Binomial
from statsmodels.genmod.families.links import logit
lr = smf.glm(
    formula="buyer_yes ~ message + age + gender + income + ...",
    family=Binomial(link=logit()),
    data=pentathlon_nptb.query("training == 1"),
    freq_weights=pentathlon_nptb.query("training == 1")["cweight"],
).fit(cov_type="HC1")
lr.summary()
```