MSPA PREDICT 450-DL-55 LEC

Micro 3: Campaign Response Modeling

Darryl Buswell

Data Preparation

```
# Load the dataset
load("data/XYZ_complete_customer_data_frame.RData")
df_cust.raw = complete.customer.data.frame

# Data structure/stats
summary(df_cust.raw)
str(df_cust.raw)
dim(df_cust.raw)
```

1 a) How many variables of each type are in Miller's integrated data frame?

The below table shows the count of each column type for the raw data frame:

```
temp <- sapply(df_cust.raw, typeof)
table(temp)
```

```
## temp
## character    double   integer
##       345       161        48
```

```
rm(temp)
```

1 b) How many variables of each type are there in the data for those customers who were mailed to in the 16th campaign?

Those customers who were mailed in the 16th campaign are identified by a flag within the ANY_MAIL_16 column. This is a subset by row rather than column. Therefore the amount of variables remains the same.

```
df_cust.16 <- df_cust.raw[df_cust.raw$ANY_MAIL_16 == 1,]

temp <- sapply(df_cust.16, typeof)
#table(temp)

rm(temp)
```

2 How many customers were targeted in the 16th campaign?

We can see that the ANY_MAIL_16 subset discussed above has reduced the number of observations from 30,779 to 14,922. Therefore, 14,922 customers were targeted in the 16th campaign.

```
#nrow(df_cust.raw) # 30779
#nrow(df_cust.16) # 14922
```

3 If you explore the data a little, you'll see that some of the character variables include levels that are blanks. See, for example, the variable named ZSKIING. How do you think you should interpret the blanks?

It seems that ZSKIING has three levels: "", "U" and "Y". This field asks whether the customer has an interest in skiing. As such, we could interpret the "Y" to be a positive response and "U" to be a negative or unknown response.

Either option could be reasonable according to the ratio of total “Y” to “U” responses. The blank responses could be data errors, refer to records where the customer did not provide a “Y” response, or refer to records where a “U” response was not recorded in-place of an unknown response.

```
#table(df_cust.raw[, "ZSKIING"])

#sum(df_cust.raw[, "ZSKIING"] == "") # 2631
#sum(df_cust.16[, "ZSKIING"] == "") # 1225
```

4 What information from XYZ or elsewhere will you use to estimate the amount that a targeted customer will spend if she or he response to a campaign mailing?

The dataset includes a number of fields labelled TOTAMTxx where the xx suffix represents the campaign number. Based on a review of “miller_xyz_data_integration.r”, it would seem that these fields relate to a total amount spent by each customer during each the 16 campaigns. For example, TOTAMT16 representing total amount spent during the 16th campaign. A simple approach would be to compare mean values for these fields for customers who were/were not part of a particular campaign mailing.

```
temp <- sapply(df_cust.raw, typeof)
#temp[temp == "double"]

rm(temp)

#df_cust.raw[ , "TOTAMT"]
#df_cust.raw[ , "TOTAMT16"]
```

5 a) You know which XYZ customers were targeted in the 16th campaign, and which ones weren’t targeted. How many of those who weren’t targeted by this campaign made purchases during the campaign?

We first subset the data frame according to those who were not mailed during the 16th campaign (ANY_MAIL_16 == 0), and then further subset this data frame according to those which made a purchase during the campaign (TOTAMT16 > 0). From this, we see that 554 customers weren’t targeted, yet made purchased during the 16th campaign.

```
df_cust.ne16 <- df_cust.raw[df_cust.raw$ANY_MAIL_16 == 0,]
#length(df_cust.ne16[df_cust.ne16$TOTAMT16 > 0,]) # 554
```

5 b) It’s probably the case that some of the customers targeted in the 16th campaign would have made purchases anyway. Is there a way you can take this “base level” purchasing into account when evaluating the effectiveness (in terms of response rate) of the 16th campaign?

We will need to ensure that any model we fit to predict response rate includes features which adequately account for any ‘base level’ purchase. This shouldn’t be an issue since the dataset allows us isolate the amount spent by customers prior to the 16th campaign using the TOTAMTxx variables. Alternatively, if we wanted to exclusively focus on the effectiveness of the 16th campaign, we could look to use some form of dummy variable for total spend over 16th campaign observations, and assess the the dummy variable effect in-sample.

6 There are a lot of variables that might be used to predict response to the 16th campaign, and some that wouldn’t make sense. How will you go about selecting variables to try out as predictors?

With such a high dimensional dataset, it would likely prove difficult to employ ‘manual’ methods for model specification. Fortunately, the power of R allows us to directly assess variable importance for a large pool of variables using a variety of model based approaches. For example, some modelling techniques provide variable importance metrics by default (e.g. boosted tree models using the ‘gbm’ package, or random forest models using the ‘RandomForest’ package). While there are also functions which exclusively focus on providing variable importance metrics, such as the ‘varImp’ function as part of the ‘party’ package, which provides a measure based on ‘mean decrease in accuracy’.