Case Study 2

Sample Report

New Start Chiropractic (NSC) has the goal of building a predictive model with two purposes in mind, to determine what factors influence the number of visits an individual has and to predict the number of visits in the future. The data in the form provided is not ready to use for predictive modeling and so needs to be modified.

The target variable is the number of visits an individual has in 2018, but this was not provided. Instead, we have each visit from 2015 to 2018. Each visit is coded with a unique ID number. Using this, we restrict the data to be from 2018 only and count the number of visits for each unique ID number. The target variable is plotted in Figure 1 as a histogram.

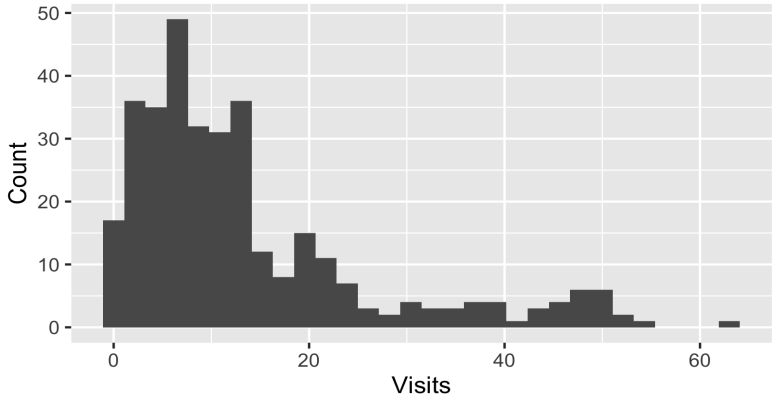

Figure 1. Histogram of Number of Chiropractic Visits in 2018

NSC has asked about using the number of visits an individual has had prior to 2018 as a predictor variable. This can be created by restricting to visits prior to 2018 and counting again by ID. One issue with this approach is that while it may be helpful for determining important factors for the number of visits, it is not useful for building a model for predictions. Consider if this model is used for the next several years. In 2020, the number of visits prior to 2018 would not be as useful. If instead the number of visits prior to the current is used in the model, the numbers would be much larger than what the model is trained on. For example, an individual who visits 10 times every year would have 30 visits prior to 2018 when the model is trained. That same individual would have 50 visits prior to 2020. So that is not a good solution. Instead, we advocate the use of the number of visits in just the year prior to the current year. In this case, it would be 2017. The variable is created as the number of 2017 visits.

There are other visit-specific variables. Specifically, the type of visit, whether or not an individual rescheduled, and how they paid. Both the type of visit and if they rescheduled are binary, and so these are aggregated by ID into new variables that are the percentage of visits that are extended versus normal visits and the percent of visits that included rescheduling. This should allow us to test the effect of the type of appointment and if a patient reschedules on the number of visits. For now, the payment method is not used because there are several levels and this is harder to aggregate.

The patients also fill out a questionnaire when they first visit NSC. The answers to these questionnaires are recorded and linked to the specific patient ID. These variables include what year the patient started, age at the beginning, sex, why the patient is visiting, an indicator for if they have previously had surgery, an indicator for if they have arthritis, and how they heard about NSC. There are a few issues with this data that were fixed.

The variable *year* is found in both data sets. For the visits data, *year* records the year of the visit while for the questionnaire data, *year* records when the questionnaire was filled out. Whenever the same information is represented by two or more variables, it is important to make sure that the information stored in those variables

match. If they do not, it is a sign that the variables either do not mean what was initially thought or that the data is corrupted somehow. In this case, matching up the year of the first visit with the year the questionnaire was completed shows a high degree of internal consistency. Only 5 out of 340 individuals did not match, which is a small enough number that it is likely each can be naturally explained.

The variable for why the patient is visiting lists all the possible reasons, so an entry that is "pain, recovery, and functionality" means that they are in for those three reasons. In this form, however, the variable is a 15-level factor variable where there is more meaning to the variables than simply individual factors. For example, "pain" and "pain and recovery" will possibly be closely linked. To fix this, we separate this one variable into 4 binary variables, one for each of the individual causes, pain, recovery, stiffness, and functionality. When each of these variables are equal to 1, this indicates that the patient has the corresponding reason for coming in, and there could be multiple reasons listed.

There are 34 missing values for the indicator if an individual has previously had surgery. Looking carefully at these values in relation to other variables, all individuals who filled out the questionnaire in 2015 are missing this value. Since there is a systematic reason here, an option is to remove these records and just deal with patients since 2016. However, it seems that the majority of individuals have answered no to this question. In fact, only 4 out of 302 who did answer this question say yes. It seems unlikely we can learn about the difference a yes response makes with only 4 cases. This variable will be deleted.

Table 1

how_know	count
internet search	95
referral	94
radio ad	79
saw store	41
youtube ad	9
news article	6
doctor referral	5
shopping cart ad	5
business card	3
television ad	3

Lastly, the variable for how a patient hears about NSC has many factor levels. These factor levels and the number of times they occur are listed in Table 1

In order to decrease the number of factor levels this variable has, the lowest factor levels will be combined into an "Other" category. Namely, everything with 9 or fewer records is recoded to "Other." Reducing the number of factor levels will help avoid bias and overfitting in the final model.

Tables 2–5 are a summary of the variables in the final version of the dataset.

Table 2

	year	age	arthritis	counts_ 2018	counts_ 2017	perc_ext	perc_resc
Min.	2015	20.00	0.0000	1.00	1.00	0.0000	0.0000
1stQu.	2016	32.00	0.0000	5.00	5.00	0.4000	0.4000
Median	2016	39.00	0.0000	10.00	10.00	0.5000	0.5000
Mean	2016	40.72	0.2284	14.04	14.53	0.5046	0.5217
3rdQu.	2017	51.00	0.0000	19.00	19.00	0.6000	0.6250
Max.	2018	60.00	1.0000	64.00	60.00	1.0000	1.0000

Table 3

sex	
Female	94
Male	96
Other	7

Table 4

how_know	
internet search	61
radio ad	49
referral	54
saw store	16
Other	17

Table 5

	pain	stiffness	recovery	functionality
FALSE	98	90	98	91
TRUE	99	107	99	106