case_2.2

April 16, 2020

1 Case 2.2

1.1 How do users engage with a mobile app for automobiles?

_"It is important to understand what you can do before you learn how to measure how well you seem to have done it." – J. Tukey

As we saw in the previous case, careful data vizualization (DV) can guide or even replace formal statistical analysis and model building. Here, we'll continue with more complex and computationally-intensive visualizations.

1.2 Introduction

Business Context. A recent trend among car manufacturers is to provide continued support through mobile applications. Features of these apps include services like remote ignition, GPS location, anti-theft mechanisms, maintenance reminders, and promotion pushes. Manufacturers are keen to maximize engagement with their app because they believe this increases relationship depth and brand loyalty with the customer. However, app usage is often limited, with many customers abandoning the app after only a short time period or never even opening it in the first place.

You are a data scientist for a large luxury automobile company. Your company wants you to uncover behavioral patterns of the users who engage with the app. They believe that if you can find discernible patterns, your company can leverage those insights to give users incentives to use the app more frequently.

Business Problem. Your employer would like you to answer the following: **"How do users currently engage with your mobile app and how has that engagement changed over time?"**

Analytical Context. In this case, we will look at data on a subset of 105 customers (out of 1,000 total app users) for the first four weeks after installing the app. This small subset of the data is chosen as a representative sample. Data were collected as part off a beta version of the app.

We will not just present a catalog of different visualizations but rather, we will look at how domain questions can guide visualizations and how carefully constructed visualizations can generate new questions and insights.

1.3 First look at the data

As always, let's begin by having a look at the data and computing a few summary statistics. The data set contains 105 rows and 116 columns. Most of the columns represent app data collected on day j ($1 \leq j \leq 28$):

Variable name	Description	Values
age	Ordinal age, coded: 1 ($<= 25$), 2 (26-34), 3 (35-50), 4 (50+)	Int: 1-4
sex	Categorical sex	Char: F, M
device_type	Android or OS X	String: Andr, X
vehicle_class	Luxury or standard vehicle	String: Lx, Std
p_views_j, j=1,...,28	Ordinal page views on day j	Int: 1-5
major_p_type_j, j=1,...,28	Majority page type	String: Main, Pr
engagement_time_j, j=1,...,28	Ordinal engagement time per day	Int: 0-5
drive_j, j=1,...,28	Indicator that user drove	Int: 0, 1

We see that a lot of the data are **ordinal variables**. An ordinal variable is a categorical variable where the categories are numbers and the relative values of those numbers matter; however, the absolute values of those numbers does not. In other words, for a given ordinal variable x, a larger numbered category means "more of x" than a smaller numbered category; however, the category number does not indicate the actual amount of x. For example, here `age` is coded as an ordinal variable; the categorical value of 3 clearly indicates "more age" than the categorical value of 1 (35 - 50 years of age vs. under 25 years of age), but the specific category value 3 or 1 is meaningless.

Below is some more information about some of the other variables:

1. The only allowable mobile platforms are Android (coded `Andr`) or OS X (coded `X`) and this is collected automatically when the app is installed; thus, we expect this variable to have no missing values.
2. The vehicle identification number was required to sign in and from this `vehicle_class` was automatically populated; thus, we also expect this variable to have no missing values.
3. The variable `major_p_type_j` is the majority page type for the user on day j. In other words, it's the type of page which is viewed most often. It's coded as a categorical variable taking the values `Main` for maintenance, `Prom` for promotions, and `Serv` for services. Here, services means the app's services (e.g. automatic start, GPS location, etc.), rather than, say, scheduling an appointment to get the car serviced (which would be categorized as maintenance).

Furthermore, a lot of the data here is "opt-in" only; that is, it is only recorded if the user was active on the app that day, and missing otherwise. For example, `p_views_j`, `major_p_type_j`, `engagement_time_j`, and `drive_j` are all "opt-in" variables.

1.3.1 Exercise 1:

What is the significance of the variables mentioned above being opt-in? What insights can we derive from this?

Answer. Because the variables listed above only record information if the user was using the app that day, the data we actually collect may not be representative of the entire population of car

owners. So we need to be careful about applying any conclusions we derive from this data to car owners in general.

Nevertheless, this pattern of missingness is still extremely useful. In particular, it tells us if a particular user was using the app that day. Given that our main goal here is to detect patterns in user engagement over time, this seems quite relevant.

Given this realization about the "opt-in" data, it makes sense for us to first understand patterns surrounding what data is missing.

1.4 Understanding and visualizing patterns in the missing data

As you saw in the Python cases, missing data is a staple of almost any dataset we will encounter. This one is no different. This dataset has substantial missing data, with nearly 60% of subjects missing a value for at least one column.

A useful tool to look at the structure of missing data is a **missingness plot**, which is a grid where the rows correspond to individuals and the columns correspond to the variables (so in our case, this will be a 106 x 115 grid). The (i, j)-th square of the grid is colored white if variable j was missing for subject i. A first pass at a missingness plot gives us:

1.4.1 Question:

Do you spot any patterns in the missing values here?

1.4.2 Exercise 2:

What are some things you can do with the dataset to visualize the missing data better?

Answer. As we have discussed before, it is critical to leverage domain expertise and our understanding of the problem in the data visualization process. Given that we believe the "opt-in" data will give us insight into app engagement, it makes sense to group the "opt-in" data together by type of information conveyed (i.e. put all the page-view variables together, all the engagement variables together, etc).

In light of this, let's remake the missingness plot with the similar variables grouped together:

1.4.3 Exercise 3:

What patterns do you notice here? Do these patterns make sense based on your understanding of the problem?

Answer. The most striking pattern here is that once a particular variable related to usage goes missing (i.e. the user stops using the app) all subsequent days are also missing that particular variable. This makes complete sense, since a common behavioral pattern with a new app is to use it intensely for a short time and then to delete it or otherwise coompletely stop engaging with it.

We can make the pattern from Exercise 2 even more apparent by not just grouping the "opt-in" data together by type of information conveyed, but by grouping them all together, regardless of type. In this case the missingness plot looks like:

1.4.4   Exercise 4:

A natural question to ask is 'what percentage of users were still engaged as of a certain day?'. How can we modify the above plot to beter visualize this?

Answer. The current version has the rows in no particular order – they're sorted according to the value of a randomly assigned user ID. If we instead order the plot according to the degree of missingness, we can easily visualize the ratio of black vs. white for any given day as a proxy for the level of app engagement:

From this plot it is immediately apparent that some subjects are dropping off and not returning; the data shows a **nearly monotone missingness pattern** which is useful for weighting and multiple imputation schemes (such methods are discussed in future cases on data wrangling). Furthermore, a significant proportion of users were engaged with the app throughout the entire 4-week period.

We now see the power of using contextual knowledge of the problem and dataset itself in the data visualization process. **The preceding four plots all contained the same underlying information, yet the later plots were clearly much easier to draw insights from than the earlier ones.**

1.5   Investigating in-app behavior

Now that we've gleaned basic insights into whether or not users engage with the app at all, it's time to do a more detailed analysis of their behavior within the app. We'll start by looking at page views.

1.5.1   Evaluating patterns in page views

To stakeholders, page views are a key measure of engagement. Let's identify patterns in the number of page views per day. Recall that page views is an ordinal variable (ordered categorical variable) coded 1-5. Here 1 codes 0-1 actual page views, with 1 indicating that the app was opened and then closed without navigating past the splash page. For each person, we have a sequence of up to 28 observations. Let's first create a parallel coordinates plot with one line per subject:

The preceding plot is extremely difficult to read. But we don't care so much about patterns for any individual user as much as the aggregate set of users. Thus, let's graph a line representing the average page views per person. The following plot shows this in black:

1.5.2   Exercise 5:

There seems to be some kind of periodicity in the above smoothed plot. What might explain this pattern?

Answer. There appears to be a period of around seven days which exactly tracks the length of a single week. This makes sense since users may drive more on the weekends, which would likely be linked to increasing app usage.

Clustering by user cohorts   Domain experts who have run qualitative studies of user behavior believe that there are different groups, or **cohorts**, of users, where the users within a single cohort behave similarly. They believe that page view behavior would be more homogeneous within any given cohort. However, these cohorts are not directly observable.

Using clustering methods (which you will learn about in future cases), we have segregated the users into three groups based on their similarities:

1.5.3   Exercise 6:

Describe the page view behaviors within each cohort.

Answer. These three plots illustrate three seemingly different usage patterns. The first appears to fluctuate somewhat around page views category `3` with no discernible pattern. The second group shows an obvious spike followed by a dramatic drop-off in page views that is commonly associated with initial app usage and subsequent disengagement. The third group appears to be regular users that obey the weekly periodicity pattern we discovered earlier quite strongly.

1.5.4   Exercise 7:

Which cohort of users do you think are more likely to look at promotional pages (major page type category `Prom`)?

Answer. If you guessed the third group you'd be right! But how can we see this from the plots above? We can't.

However, from talking to our colleagues, we learn that promotions are pushed out every Friday which causes a spike in some users who tend to peruse them over the weekend. Again, we see that input from domain experts is critical here.

1.5.5   Analyzing patterns in major page type

Let's have a look at the major page type over time across our three user cohorts:

From this, we can see that the third group is indeed the most engaged with the promotional pages.

1.5.6   Exercise 8:

What are some potential next steps if you wanted to do a deep dive into user page view behavior? What additional data might you want to collect on users?

Answer. We might want to collect data on engagement behavior with specific promotional materials or more detailed information about usage of specific service apps. For example, if the user

remotely starts their car on weekday mornings, we might want to push out reminders like "Would you like me to start the car?" at a given time. To study disengagement behavior, we might conduct user surveys on focus groups to learn why they discontinued with the app.

1.6 Predicting dropout from page view behavior

Because page view behavior is believed to be strongly related to engagement with the app and likelihood of discontinuation, we would like to see if we can predict the point of disengagement by analyzing the page view behavior within each cohort. We start by simply labeling the last observation (i.e. day of usage) for each subject with a large red dot:

1.6.1 Exercise 9:

Do you notice any patterns in page views preceding dropout?

Answer. Within groups 2 and 3, low page views seem to precede dropout, but there does not appear to be any such pattern in group 1.

1.6.2 Exercise 10:

Work with a partner. Based on the preceding visualizations, propose an adaptive intervention strategy that monitors a user's page views and then offers them an incentive to continue using the app right when we believe that the incentive would have the most impact. Assume that you can offer at most one such incentive during the first four weeks of app use.

Answer. As the page view behaviors and disengagement patterns vary dramatically across user types, we should adapt our intervention strategies to these patterns.

For the first group, absent any additional information, we might offer the incentive at the two-week mark when disengagement seems to accelerate.

For the second group, we might try to estimate the spike trend in app usage for each user and offer the incentive right when we expect their peak engagement. This might help prevent the steep dropout following peak engagement.

For the third group, we might offer the incentive during the middle of some week – perhaps just before the lowest point of engagement when dropout risk is highest. Of course, selecting the exact week to deliver this incentive is more challenging – we would need more data for that.

Other points to consider:

1. Above, we have focused on "catching" the user before they disengage; an alternative would be to send them an incentive after they drop out.
2. Focus groups or information from domain experts can provide key insights about the timing and types of interventions. For the third type of user it makes sense to focus on promotions, but for other users it may make sense to offer maintenance discounts or other services.
3. The interventions may also need to vary by age and sex or other features.

1.7 Conclusions

We explored usage and disengagement patterns among users of a mobile app for a car manufacturer. We saw that most users still remained engaged with the app even after 28 days, and that there were three significantly distinct cohorts of users. We used these patterns to generate ideas for intervention strategies that might be used to increase app usage and reduce disengagement. These visualizations are an excellent starting point for building statistical models or designing experiments to test theories about drivers of disengagement.

1.8 Takeaways

In this case, you looked at more types of plots and how to draw conclusions from them. You also learned how these conclusions can drive further questions and plotting. Some key insights include:

1. Sometimes it is important to reorder the data according to some variable in order to derive insights (as we saw with the missingness plot).
2. Sometimes additional computation or data manipulation is required in order to tease a meaningful pattern from a data visualization (as we saw with the clustering & averaging for the parallel coordinates plots with the three cohorts).
3. Domain knoweldge and understanding the context of the problem and data at hand is crucial. Without this, we would never have been able to create the visualizations we did and draw the conclusions we did from the missingness plot and the parallel coordinates plots.