Thinkful Data Science Bootcamp

Unit 1, Lesson 1, Drill

10/19/2017

Assignment

As we outlined in the previous section, a lot of data science work can be classified as translation: taking a question and translating it into mathematical or statistical tests or taking statistical results and translating them into something everyone can understand.

Below we have a series of questions for you to translate into a technical plan. For each question, describe how you would make it testable and translate it from a general question into something statistically rigorous. Write your answers down in a shareable document and submit the link below.

1. You work at an e-commerce company that sells three goods: widgets, doodads, and fizzbangs. The head of advertising asks you which they should feature in their new advertising campaign. You have data on individual visitors' sessions ([activity on a website](https://en.wikipedia.org/wiki/Session_%28web_analytics%29), [pageviews](https://en.wikipedia.org/wiki/Page_view), and purchases, as well as if those users [converted](https://en.wikipedia.org/wiki/Conversion_marketing) from an advertisement for that session. You also have the cost and price information for the goods.
   * Define goal
     1. For example, are we trying to improve final sales, overall profitability or advertising ROI?
     2. Assume we are trying to maximize the total profit.
   * For each product determine the possible increase in sales (given that price and cost are fixed) from an increase in advertising for that product
     1. Proxy advertising by a combination of characteristics of visitors’ sessions
     2. Identify a basic relationship (e.g. linear) between advertising changes and sales of a particular product
   * Allocate advertising budget proportional to increase in sales due to advertising increases times profit per unit.
2. You work at a web design company that offers to build websites for clients. Signups have slowed, and you are tasked with finding out why. The [onboarding funnel](https://en.wikipedia.org/wiki/Funnel_analysis) has three steps: email and password signup, plan choice, and payment. On a user level you have information on what steps they have completed as well as timestamps for all of those events for the past 3 years. You also have information on [marketing spend](https://en.wikipedia.org/wiki/Marketing_spending) on a weekly level.
   * In this case we cannot find out with complete certainty why signups have slowed, but we can try to formulate a few simple (& reasonable) hypotheses and see if we can reject any of them or find evidence that they may have happened (the latter does not mean the hypothesis is true).
   * Remark: it is unclear to me whether “signups have slowed” refers to just the first part of the funneling process or to all of them. I will assume it refers to all of them.
   * Formulate a few hypotheses and tests
     1. Signups have slowed because people don’t see our product enough (i.e. advertising has slowed)
        + Check: is there a relationship between the changes in weekly advertising and sign-ups, with appropriate lag. The lag should be the typical amount of time it takes to go from email login to plan payment.
     2. Signups have slowed after we increased our product offering (by creating more options)
        + Check before and after increase whether the time to payment has increased in a significant fashion (relative to pre-change period). For example, check typical changes before the change and see how many orders of magnitude the fluctuation after the change is.
     3. Assuming no change in offering, signups have slowed for reasons unrelated to our plans, irrespective of our changes in advertising.
        + Check: controlling for changes in advertising, the time to payment has increased across the board. One can use a linear regression approach to control for changes in advertising.
3. You work at a hotel website and currently the website ranks search results by price. For simplicity's sake, let's say it's a website for one city with 100 hotels. You are tasked with proposing a better ranking system. You have session information, price information for the hotels, and whether each hotel is currently available.
   * Identify the goal
     1. Depending on how the website makes an income, we need to define “better ranking system”.
     2. One reasonable assumption is that “better” is related to increasing reservations (number or $ amount) made through the website.
   * Diagnose the problem
     1. Are there missed opportunities? For instance, given a period of time of relevance to us:
        + Are there many customers (e.g. large percentage) which visit our website, but do not try to book at all?
        + Are there many customers which visit, but take a while to book?
        + Are there many customers which visit, try to book, but find that the rooms are not available?
        + We can compare these numbers across time and with competitors (if they are available) to get an idea of the evolution of the problem.
   * Depending on what the opportunities are, adopt a possible solution
     1. If first case, we assume customers just don’t find what they are looking for so we try to develop a few other metrics about the hotels besides price. For instance, we could try to add quality or distance fields for hotels.
     2. If second case, these are probably bargain hunters and price is the most important piece of information.
     3. If third case, construct a ranking based on price and availability rather than just price.
4. You work at a social network, and the management is worried about [churn](https://en.wikipedia.org/wiki/Churn_rate) (users stopping using the product). You are tasked with finding out if their churn is atypical. You have three years of data for users with an entry for every time they've logged in, including the timestamp and length of session.
   * Define usage of product
     1. For a given period define the usage as the length of session during that period divided by the length of the period.
     2. For simplicity, assume the length of the period of time we are concerned with is 3 months.
   * Define “stopping using the product”
     1. Define stop using the product as the bottom 5-10 percentile of the distribution of usage of product for a given period of 3 months (the percent should be related to other measures of interest like profitability of advertising for that category). This number has units *hours/3months* (e.g. 0.5hours/3 months)
   * Define atypical churn
     1. Compute the number of people that have stopped using the product in a period of 3 months over the first 2 years (or over the sample of 3 years minus period of interest).
     2. From the distribution of these numbers, select the 90th percentile as typical churn.
   * Are we dealing with atypical churn?
     1. If the current period number is higher than the 90th percentile defined above then we can say that we have atypical churn.
   * An alternative way of doing this is to fit a distribution to the empirical number of people using the product in a period of 3 months and use that.