MSSP Portfolio Report

Xuan Zhu

Department of Statistics

Boston University

Table of Contents

Abstract

This portfolio includes five major projects that I have participated in during my graduate study. The Shire Project (SP) and the Boston Public Schools project (BPS) are known as the “Partner Projects”, in which our team worked with industry partners to solve real business world problems. In SP, the team was divided into three components with different focus, and I was in the group which mainly measures the Market and Payer Effects of the company’s new drug on market. In BPS, we did enrollment projection to help the system save budget and I predicted the number of the enrollment of kindergarten students in 2019-2020 academic year across the Greater Boston Area. SP was conducted under the guidance of our professor Haviland Wright and BPS was of Masanao Yajima. Meanwhile, the living consulting projects, the Sleep & Cortisol Project (SC) and the Athletic Leadership Program (ALP), were supervised by Nathan Joseph. I worked with Chaoqun Yin, Yudi Mao, Douglas Goon and Andrew Zhang to offer statistical support to our school students and faculties such as survey analysis and generalized linear model analysis. It needs to be noted that not all consulting projects are included in this portfolio. The incomplete one, the Neuron Project and another one without specific data are excluded after careful consideration. Last but not least, the Web Analytics Project is cooperated with the FoodEasygo Co. as part of the CS688 course requirement, which I think helped me polish the skills of data mining and data analysis a lot.

The projects are presented without a certain order, and all sensitive data is blurred or left out according to the confidentiality requirement.

Shire Project

I. Introduction

Our industry partner, Shire Plc, is a global specialty biopharmaceutical company. In Aug 2016, Shire's new medication Xiidra came into the Dry Eye Disease (DED) market to compete with the current market giant, Restasis. By the end of this project, Xiidra and Restasis are the only two FDA approved medications that can treat dry eye syndrome by prescriptions, and Xiidra launch has already become a massive success that not only keeps grabbing the market share from Restasis but also increases the whole size of market. However, the increasing trend was plateauing in the late 2017. Our partner is interested in:

- The insights behind their current and history sales data. E.g. What affects a healthcare professional to prescribe Xiidra over Restasis? And what may be the reasons behind the decreasing increasing trend?
- Whether there is a way to measure the Payer Effects & Sales Call Effects
- Future prediction

II. Data and Methods

Data source is multiple. We have the sales call records from four different sales call teams and all the prescription information from Healthcare Professionals (HCP) registered with a unique Shire ID. Each HCP has records of personal address, decile, the day of visiting, etc. Each patient is identified with a unique ID, and is classified into two types. If one patient had no records of using Restasis before but directly used Xiidra after he/she got DED, he/she was classified as "new patient" to the Shire. Otherwise, the prescriptions were recorded in the "switch" datafile. Some patients may insist on using Restasis, or switch to use Xiidra at some point. The main reason that drove patients to choose medication is the insurance plans that cover patients, and we call it "the Payer Effects".

In this project, our team divided the problem into three components:

- Data Wrangling and The Exploratory Data Analysis. We cleaned, re-structured and enriched source data into a desired format and visualized the competition of Xiidra vs. Restasis. SQLlite and Tidyverse/ggplot2 packages of R are main tools applied.
- Market and Payer Effects & Sales Call Effects Analysis. We examined the effects across states, times, different HCP groups, etc.
- Modeling. We also tried to use the logistic model to regress the probability of a new patient choosing Xiidra over Restasis on five predictors: States, Insurance Plan, Time, Sales Calls and HCP Decile.

III. Results & Discussion

Even though our model has 72% classification accuracy over historical data, it is not appropriate for future prediction in the current form. Responses to sales calls vary by HCP, which means that we need multilevel models to account for structure of clustering within HCP, states and payers. However, due to time and hardware limitation, multilevel models are not possible for us to perform out because we have over thirty thousand HCP levels with millions of prescriptions.

For the Sales Call Effects, we conformed that HCP with higher decile received more sales calls and they are more willing to prescribe Xiidra over Restasis, and we also discovered that calls to HCPs with large Government Payer Mix may be less effective than those with Non-Government Payers. Since observations are not independent, we suggest Shire to use randomized control experiments (A/B testing) to glean effects of varying sales call volume on different HCP groups.

IV. Conclusion

What factors influenced writing of Xiidra prescriptions based on historical data? Through data analysis, we concluded that:

- New patients insured by non-government plans typically received Xiidra prescriptions at higher proportions compared to new patients on government plans

  -This relationship holds up even when examined across different insurers, dates, or sales call volumes
- For prescriptions to new patients, the proportions of Xiidra versus Restasis differ significantly across Books of Business (BOB) and for each BOB, these proportions vary over time
- Increased sales call volume is associated with higher proportions of Xiidra prescriptions when examined on HCP or decile level
- The higher an HCP's patient government payer mix, the lower the proportion Xiidra prescriptions for that HCP
  - Physicians with very high government payer mix still received significant sales call volume

Sleep & Cortisol Project

I. Introduction

The consulting project Sleep & Cortisol is based on a biological experiment which is to investigate the relationship between infants' sleep quality and cortisol levels. Cortisol is the human's body main stress hormone, working with certain parts of the brain to realize multiple functions, such as regulating blood pressure and controlling the sleep cycle. Our client Charu Tuladhar who initiated this project is a PhD student from the Psychology and Brain Sciences Department of Boston University. Her interest is focused on how cortisol levels deal with the sleep cycle of infants around 12 months old. The goal of our team is to provide the client with a statistical analysis on her experiment result. Within the consulting period, we have achieved three main objectives:

(a) We produced an analysis plan.

(b) We carefully examined & restructured the source data, and then performed Exploratory Data Analysis (EDA).

(c) Based on the results of EDA, we fit a linear mixed model.

II. Data and Methods

The experiment recorded ~90 infants in a two-week period. For the study, sleep onset and sleep duration were measured using actigraphy. Sleep onset is the time when infants fell asleep, and sleep duration is defined as the time between sleep onset and wake time. On the following day, parents collected saliva from the infants in the morning, afternoon, and night. The measurements were timestamped, and then collected and frozen by the client before analyzing and determining cortisol levels. The three levels were used to calculate cortisol level slope, as well as area under the curve (AUC). Each infant had three days of record within the two-week window, but they are not all on consecutive days. Furthermore, the client has data on age, gender, socioeconomic status, sleeping routing according to the parents, height, and weight for each infant.

There are two portions of analysis our client is looking for from the experiment results. The first analysis is the relation between infants' cortisol level as response and their sleep onset and duration time as predictors. The client expects that a later bedtime will yield a flatter slope and greater AUC, and longer sleep duration will yield a steeper slope and lower AUC. The second analysis is to predict the sleep onset time of infants using the cortisol level recorded that night as the predictor. The client hypothesizes that higher cortisol levels at night will lead to later sleep onset.

III. Results & Discussion

- Analysis 1

The overall trend between AUC and Sleep Duration/Onset roughly conforms the client's hypothesis. However, in the plot of AUC and Sleep Duration, the smooth line fluctuates a little where sleep duration is between 9am and 10am. In the plot of AUC and Sleep Onset, the right tail goes down, which is opposite to the overall trend. That would be the result of few observations where sleep onset time is between 12am to 2am. The curve of the slope fluctuates overall, so we are

not able to say if they also agree with the hypothesis. We also visualized basic statistics of descriptive variables, though we did not see significant difference among groups.

- Analysis 2

With the general raw data points, there is clearly a positive trend between Bedtime and Bedtime Cortisol levels. A high chaos score, which indicates a more chaotic household, has a higher slope, or a later bedtime with a higher cortisol level. Age, ITN, Job Score, and Education Score have an inverse effect, meaning things like an increased age has an inverse slope relationship, where a lower cortisol level is associated with a later sleep onset.

IV. Conclusion

For this project, we were able to successfully separate tidy datasets for each of the analyses. In terms of the EDA, we can see that the data does not behave as we expected, given that there is severe heterogeneity in analysis 1. This could be attributed to the fact that we are working with already winsorized and z-score transformed variables. It would be better if we could perform EDA with the raw data, allowing us to understand which methods were applied to the data. Unfortunately, we do not have access to the raw data.

We are able to conclude that there are positive relationships between sleep onset and bedtime cortisol, as seen in the plot generated in analysis 2. In addition, in terms of demographic variables, chaos score also seems to play a positive role in the relationship between sleep onset and bedtime cortisol. Although we only identified one variable as being helpful, the inverse relationships between the slope of sleep onset and bedtime 8 cortisol and demographic variables such as age, income to need ratio, job score, and education score, are relationships that are worth looking into. Mixed modeling may not be the best choice for this project, as we discussed before. The number of observations is too limited for mixed model. In addition, from analysis 1, we can see that the data is not homogenous enough for a linear mixed model. We may try other transformations first and then apply linear models, however, this is all speculative without the raw data.

Web Analytics Project

I. Introduction

FoodEasygo Company is an online food ordering and delivery platform. It was founded in 2014 and the current version of the website was launched in 2016. Our team worked with FoodEasygo for the Google Analytics (GA) project as part of CS 688 requirements. In a 7-week period, the team interviewed the owner of the company, evaluated the design of the website, tracked website visiting activities with GA, analyzed user behavior, and provided suggestions to revise their website based on our observations. Basically, we applied A/B testing, web mining and analytics skills in this project.

The company is a new start-up in Boston so it had never used Google Analytics before. One of the initial motivations for cooperating with us is to have a better understanding of their customers. By analyzing Google Analytics results, we can provide them insights on demographic information of their customers as well as on their behavior.

II. Data and Methods

All the data was automatically gathered by Google Analytics. The dataset was then generated as one-month continuous website traffic data in October, 2018. We analyzed the data from the following perspectives:

- The web design assessment. We assessed the pros and cons of the website from a potential customer's perspective and gave suggestions.
- The visibility of the Site. The owner of the website mentioned they are trying to promote the website by posting their ad on Dealmoon and Facebook. They also have an official WeChat account and post some deals on that. By inserting the tracking code, our team was able to track which ads the visitors are clicking from and then drew out the traffic map.
- Visitor's behavior and patterns. The main goal is here is to answer the question of whether the website is attracting its target customers.
- Measure of Success. We implemented two measures of success: the conversion rate and a successful A/B testing on their new lunch service page. The conversion rate is the percentage of website visitors who finally pay the service on the site.

III. Results & Discussion

From the traffic treemap, we can tell that most users search their website and order food directly, and some use keywords to find them. Visitors from Referral is not a significant part, thus ads posted on other sites did not attract many visitors to this website.

The target audiences of the website are college students and young working adults. As we can see from the GA results, most of the visitors to the website are between 18 to 24 years old. It means that the website is attracting their target audiences.

Around 60% of visitors are male. Therefore, the design of the website can cater more towards male preferences, and the owner can also prioritize collaboration with restaurants that are popular among males.

Athletic Leadership Program

I. Introduction

II. Data and Methods
III. Results & Discussion
IV. Conclusion

Boston Public Schools Project

I. Introduction

II. Data and Methods

III. Results & Discussion

IV. Conclusion