The Weekend Effect
November 26, 2018

INTRODUCTION

Over the course of the past 15 weeks, this class has introduced and explored a wide array of topics across the healthcare analytics domain. Broadly, these topics have been broken down into three broad themes: consisting of data understanding and preprocessing, statistical methods, and most recently model building and validation. As we have moved through the semester, students have demonstrated an acquired knowledge around a subset of methodologies pertaining to each theme, with respect to the strengths, limitations, and ultimately the manner in which they can be executed in functional python code.

Now, in our final week of class, it is time to put everything you have learned together. Throughout the next three class sessions, students (in groups of 4-6)\(^1\) will complete a case study requiring the design of an analytical solution to address the following research question:

"Are patients admitted to the intensive care unit (ICU) on a weekend more likely to die in the hospital than those admitted on a weekday?"

The question itself is one of particular interest in the health analytics field today. It has been adapted from a data challenge run at an international conference last year, sponsored by the MIT lab who maintains the MIMIC data. However, as the time required to address the full scope of such a question is far beyond that available this week, I have provided a set of specific milestones to guide your groups initial steps.

Note: I have left the specifics of each milestone intentionally vague as to allow for the most freedom possible as you design your study.

\(^1\)The same groups as Assignment 4
DATA UNDERSTANDING / DATA PREPROCESSING

In the interest of keeping this case study feasible to complete within class time, I have extracted a wide array relevant data elements from across the MIMIC database\(^2\). Although, I have only narrowed down the data, it remains up to your group to select what you believe are relevant features. Proceeding to clean and analyze it in such a manner to provide a justifiable answer to the question.

As such, the first milestone is broken into three smaller parts, and covers primary aspects of cohort identification, feature selection, and data cleaning. Details of each can be found in the respective sections to follow.

**Cohort Identification**

The first portion of this milestone will require the identification of a study population (cohort). As we have all semester, take the full set of MIMIC patients (contained in the data I provide), and define a set of constraints to focus on a specific set of patients your group believe will provide a reliable (and technically correct) analysis of our research question.

Please explicitly state the criteria used for cohort selection, and why you believe such a criteria is necessary. Use markdown cells to so. Additionally, please provide the unique number of patients remaining after each filtering step.

**Feature Selection**

Once you have identified (and justified) a study cohort, you will need to complete the selection of features to include in the final analysis, both by drawing on those features in the files I provided, as well as creating some of your own.

Note, I have intentionally provided more features than you need, or can even process in the time allotted. Some features may not be useful, others may hurt your model performance, it is up to your group to select a set that makes sense. I suggest focusing on attributes of the patient (demographics), and their admission (as to not bias the results with things that happen post-admission). While also performing some obvious feature engineering: i.e. determining the day of the week admission occurred on\(^3\). Please list the final set of features used in the analysis.

**Data Cleaning**

The final aspect of this milestone requires your group to perform at least a cursory data cleaning on the features selected. This cleaning should aim to account for aspects of the MIMIC anonymization (masked data ), missing data, small or similarly named groups for nominal features, etc.

Note: I appreciate this case study is being completed within a week in-class. Be thorough with the cleaning, as the data left will impact the quality of your results, but understand there is limited time to complete the study. I have already performed a significant amount of data cleaning for you, and addressing some of the common major issues listed above will be sufficient for this step.

\(^2\)Stored in the CaseStudy.pkl file within the CaseStudy folder in the MIMIC directory of my shared AFS space.

\(^3\)I highly suggest creating a binary flag for weekend, aggregating weekdays and weekend days together, rather than computing each day separately. In this way, the statistical comparisons are much more straightforward.
DATA EXPLORATION / STATISTICAL METHODS

The next milestone will focus on better understanding the features selected for analysis and how they present within each class (weekday vs. weekend admissions). Here your group will need to produce some summary statistics (and visualizations should you be so inclined) to compare features across patients admitted to the ICU on the weekday vs. weekend.

Taking this one-step further, use some of the statistical methods covered in Week 4 to identify at least 3 features that have a significant difference between the two groups (weekday and weekend admissions).

MODEL BUILDING / VALIDATION & INTERPRETATION

The final milestone will draw on the cohort and features selected and cleaned to construct an analytical model designed to address the overarching research question presented in this case study. The choice of model, features, reference categories, interactions, and target variable will be left at the discretion of your group.

MODEL BUILDING

In an effort to provide a little guidance, and to demonstrate the weekend effect is (crudely) present in the MIMIC data, I have created an example model on a simple cohort that has been minimally filtered (and not cleaned at all). The model is unadjusted (i.e. a model which includes only the factor of interest), where the feature corresponds to weekend admission and the outcome variable of mortality during a hospital admission. Results for this model can be seen below in Figure 0.1.

|            | Coef. | Std.Err. | z     | P>|z| | [0.025] | [0.975] |
|------------|-------|----------|-------|-------|--------|--------|
| Intercept  | -2.1127 | 0.0185  | -114.2000 | 0.0000 | -2.1490 | -2.0765 |
| C(inday_icu_wkd)[T.weekend] | 0.2992 | 0.0368 | 8.1288 | 0.0000 | 0.2270 | 0.3713 |

Figure 0.1: Model Formula: ‘hospital_expire_flag ~ C(inday_icu_wkd)’

Remember the coefficients are on log scale, so when we exponentiate it, we get the odds-ratio: \( \exp(0.2992) = 1.35 \). Therefore, looking at these crude rates and odds ratios, we can see that patients admitted on a weekend have about a 35% increase in the odds of dying in the hospital when compared to those on a weekday. This effect is statistically significant (p< 0.001).

Great! So you’re done right? Copy and paste my model and move on...

Not quite. This is only a start, there are likely some confounding effects that must be controlled for using the features you selected, and maybe even some interactional effects between their levels and weekend admission that can better elucidate the manifestation of the weekend effect.

\(^4\)The groupby function will be your friend here

\(^5\)hospital_expire_flag (0: Discharged, 1: Died) – inday_icu_wkd (Reference-level 0: Weekday admission)
**Validation & Interpretation**

Explore this feature space and come up a final model (or set of models) your group believes best addresses the question of if weekend intensive care unit admissions have a higher probability of mortality than weekday admissions. Justify your model specification decisions (features included, reference levels used, interactions).

Finally, no study is without its limitations. Spend a few minutes and discuss the limitations of your groups approach. This can be included as a single markdown paragraph under the heading Limitations. Here consider adding some discussion around the features and models you selected, the MIMIC data used, and even the definition of the question itself.

**Deliverables**

Only a single Jupyter notebook (containing the names of all group members) is required for each group. Downloaded as HTML, the notebook must contain the following elements

- An explicit review of the filtering criteria to obtain the final study cohort (listing the steps and justifying why they are needed).
- A list of the final feature set used in the analysis.
- Some basic summary statistics for the features selected, and a clear demarcation of three which have statistical differences between weekday and weekend admission patients.
- A model (or set of models) results that rigorously address.
- A review of the design decisions made when constructing the model, and a brief discussion acknowledging the limitations of the work done during the case study.

**Grading**

- Milestones – 85 Points
  - Data Understanding / Data Preprocessing – 30 Points
    * Was the group able to identify / justify a sensible cohort?
    * Were appropriate features selected for analysis?
    * Was reasonable attempt made to prepare the data for analysis?
  - Data Exploration / Statistical Methods – 15 points
    * Were the statistical methods appropriate for the data types being evaluated?
  - Model Building / Validation & Interpretation – 40 Points
    * Do the model results support a conclusion around the weekend effect?
    * Was consideration given to model specification decisions and study limitations?
- Technical Correctness – 10 points
  - Are the methods used appropriate for the type of data and question being answered?
- Critical Thinking and Creativity – 5 points
  - Do the solutions provided to each milestone show an understanding of the concepts they implement? (I.e. Are they simply just copied notebook code, or do groups have an understanding of the design decisions made at each step).
I hope that at this point, the (adjusted) model your group constructed has yielded some insight into at least one population of patients for whom weekend admission appears to have a significant relation to increased mortality rates. With the time left after completing the three milestones, dig into these patients.

Using the full breadth of the MIMIC database\textsuperscript{6}, can you identify any differences in the data of those patients admitted on the weekday and weekend.

\footnote{\textsuperscript{6}Found in the original MIMIC folder on my shared AFS space}