36-618: Experimental Design & Time Series

Midterm Project Overview

Spring 2018

Components Due at Various Times

Introduction

As a data scientist, it is important to not only display strong analytical skills, but also interpretive skills as well. Building predictive models and providing descriptive analyses is only half the battle — identifying insights, communicating results, and relaying actionable recommendations is what differentiates a data scientist from a true statistical consultant. For this project you will practice implementing both the hard and soft skills necessary to liaise with a client in a professional setting. You will build predictive time series models and ultimately deliver results in both an oral and written fashion. **What story will you tell?**

Overview

You are on an accelerated path to becoming an accomplished statistician: machine learning topics abound and the anticipation of building your own models excites, but there are so many algorithms, tuning parameters, and assumptions of which to keep track. Additionally, there is only a finite amount of time each day. How are you going to survive?

The bottom line is: **You will find a way.**

It's time to unleash all that you've learned, start answering questions, and constructing solutions as though you are a true data scientist. The data is your oyster, and R/SAS/machine learning are your shucks. How in the world will you bring focus to your endeavor?

The bottom line is: **You will find a way.**

While the primary goal of supervised learning is generally focused on predictive accuracy, you will be expected to lead your audience through descriptive insights as well. Though employing time series methodology in order to make forecasts will be of utmost importance, you will aim to not only create models that predict well, but also allow yourself to describe data insights drawn from exploration. Furthermore, you will have research questions from a client that must be addressed among the broader analyses you perform. It seems like you're doing double-duty, and time is short. How will this all be possible?

The bottom line is: **You will still find a way.**

Data

The framework will be through the lens of daily FitBit data collected from your client since December, 2015. You will be provided with daily measurements on a user's body, activities, and sleep information. **You have signed an NDA and have agreed not to share this data externally.**

Teams

Because there is so much data at hand, the client has split you into specialized teams. While you are allowed to use data from each data sect, your primary focus is upon your own data realm:

Body Team #1	**Activities Team #1**	**Sleep Team #1**
Han Bao	Bishal Karki	Ryan Christianson
Robert Feldman	Melody Owen	Eva Gjekmarkaj
Arthur North	Qin Pu	Maitreyi Gunye
Steven Sreng	Andrew Resnikoff	Grace Lee
Dejia Su	Melinda Wang	Letti Liu
Leo Yoon		
Body Team #2	**Activities Team #2**	**Sleep Team #2**
Tanguy Dauphin	Thomas Goode	Yiming Gao
Rebecca Gu	Jedediah Grabman	Nolly Gibbs
Tianyi Lan	Jiaqi He	Emily Wang
Taiyan Liu	Vanessa Vidal	Junyi Zhang
Amanda Mummert	Steve Zhang	Jenny Zhu

Starting Notions for Research Questions of Interest

Body Teams

- Your client is interested in gaining weight (goal: 180lbs), but maintaining a healthy BMI.
- Muscle mass is of utmost importance, and fat mass is undesirable.
- What, if anything, is associated with gaining weight?
- What are your daily forecasts for weight in March, 2018?

Activities Teams

- Your client is interested in maintaining a consistent exercise regimen.
- The dichotomy between calories burned and sedentary minutes is of particular concern.
- What, if anything, is associated with burning calories?
- What are your daily forecasts for calories burned in March, 2018?

Sleep Teams

- Your client is interested in maintaining a consistent sleep schedule.
- Minutes asleep is of most importance; the client would like to aim for 7.5 hours per day.
- What, if anything, is associated with minutes asleep?
- What are your daily forecasts for minutes asleep in March, 2018?

All Teams

- Are there any trends? Is there any seasonality?
- Are there any inconsistencies/anomalies within the data?
- Is there any missingness?
- Is there any identifiable explanation for the above? Why do you think this is?
- What did you do to account for the above? Why did you make those choices?

Expectations

As always, preparation will be key. Successful projects will encompass a plethora of skills including, but not limited to, the following:

- Submission in respect to deadlines.
- Background knowledge of dataset(s).
- Communication of motivation: why do we care?
- Research questions of interest: what do you want find out?
- Answers to research questions: what have you uncovered?
- Demonstration of presentation skills:
    - Oral methodology.
    - Written methodology.
- Time management (not going over the allotted time).
- Ability to answer audience questions effectively and efficiently.
- Balance of complexity and simplicity.
- Explanation of future work: what would you do if given more time, data, etc.?
- Demonstration of EDA skills:
    - Numeric methodology.
    - Graphic methodology.
- Demonstration of time series modeling/machine learning skills:
    - Supervised methodology.
    - Unsupervised methodology.
- Accuracy of forecasts.
- Ability to research and implement new machine learning skills.
- Ability to assess model weaknesses and identify improvements.
- Ability to manage a team workflow.

Midterm Project Presentation Preview

- **Due uniformly on Monday, April 2 at 10:00pm.**
- In-Class presentations to take place:
    - Tuesday, April 3: Body Team #1, Activities Team #1, & Sleep Team #1
    - Thursday, April 5: Body Team #2, Activities Team #2, & Sleep Team #2
- 20 minutes, followed by a few audience questions.
- Create a presentation (e.g. PowerPoint, Keynote, .pdf, etc.) that summarizes your findings, addresses the research questions, etc.
- You will answer questions about your project from the class. Additionally, you are also expected to ask questions about others' projects.
- [More details to follow.]

Midterm Project Report Preview

- **Due uniformly on Friday, April 13 at 10:00pm.**
- An overall discussion of your project in the Executive Summary — Introduction, Methods, Results, & Discussion (ES-IMR&D) style.
- [More details to follow.]

Do your best & forget the rest.