ST 514 Final Project (due May 2, 11:59 PM)

In this project you will create a group report and a SAS program. These should both be uploaded to Moodle via the assignment link. It is sufficient for your group to upload a single version, under the Moodle account of one student.

Notes about the project:

- You'll be working in your assigned groups (A, B, C) for this project. You can discuss the project with others in the class who aren't in your group, but you cannot share code between groups.
- I recommend using a collaborative editor like google docs to write your reports. You can create a shared folder and put your SAS programs in there as well.
- The project has some open-ended elements. The goal is to take the concepts and principles we've learned so far and apply them to real data.
- Be sure that your SAS program adheres to the SAS program file submission guidelines (available on Moodle in the "Resources and Information" Section).

Your mindset for the project:

You have been hired as a data analyst by a marketing firm, primarily to predict whether a customer will accept a coupon for a nearby business, where the coupon is presented by their car's mobile recommendation device. There are other ancillary questions.

You will use a dataset from the UCI Machine Learning Repository

and described in the paper

https://jmlr.org/papers/volume18/16-003/16-003.pdf

From the author description:

" We are interested in investigating five types of coupons: bars, takeaway food restaurants, coffee houses, cheap restaurants (average expense below $20 per person), expensive restaurants (average expense between $20 to $50 per person). In the first part of the survey, we asked users to provide their demographic information and preferences. In the second part, we described 20 different driving scenarios (see examples in Appendix B) to each user along with additional context information and coupon information (see Appendix B for a full description of attributes) and asked the user if s/he will use the coupon."

Dataset:

Although Appendix B of the paper is helpful to understand the variables, the data from the UCI Repository and the variable descriptions should be considered the official source for this project.

The dataset for this project comes from the UCI machine learning repository. The particular dataset is available at

https://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation#

The following attribute information (including any misspellings) is taken from the site:

Attribute Information:

destination: No Urgent Place, Home, Work

passanger: Alone, Friend(s), Kid(s), Partner (who are the passengers in the car)

weather: Sunny, Rainy, Snowy

temperature:55, 80, 30

time: 2PM, 10AM, 6PM, 7AM, 10PM

coupon: Restaurant(<$20), Coffee House, Carry out & Take away, Bar, Restaurant($20-$50)

expiration: 1d, 2h (the coupon expires in 1 day or in 2 hours)

gender: Female, Male

age: 21, 46, 26, 31, 41, 50plus, 36, below21

maritalStatus: Unmarried partner, Single, Married partner, Divorced, Widowed

has_Children:1, 0

education: Some college - no degree, Bachelors degree, Associates degree, High School Graduate, Graduate degree (Masters or Doctorate), Some High School

occupation: Unemployed, Architecture & Engineering, Student,

Education&Training&Library, Healthcare Support,

Healthcare Practitioners & Technical, Sales & Related, Management,

Arts Design Entertainment Sports & Media, Computer & Mathematical,

Life Physical Social Science, Personal Care & Service,

Community & Social Services, Office & Administrative Support,

Construction & Extraction, Legal, Retired,

Installation Maintenance & Repair, Transportation & Material Moving,

Business & Financial, Protective Service,

Food Preparation & Serving Related, Production Occupations,

Building & Grounds Cleaning & Maintenance, Farming Fishing & Forestry

income: $37500 - $49999, $62500 - $74999, $12500 - $24999, $75000 - $87499,

$50000 - $62499, $25000 - $37499, $100000 or More, $87500 - $99999, Less than $12500

Bar: never, less1, 1~3, gt8, nan4~8 (feature meaning: how many times do you go to a bar every month?)

CoffeeHouse: never, less1, 4~8, 1~3, gt8, nan (feature meaning: how many times do you go to a coffeehouse every month?)

CarryAway:n4~8, 1~3, gt8, less1, never (feature meaning: how many times do you get take-away food every month?)

RestaurantLessThan20: 4~8, 1~3, less1, gt8, never (feature meaning: how many times do you go to a restaurant with an average expense per person of less than $20 every month?)

Restaurant20To50: 1~3, less1, never, gt8, 4~8, nan (feature meaning: how many times do you go to a restaurant with average expense per person of $20 - $50 every month?)

toCoupon_GEQ15min:0,1 (feature meaning: driving distance to the restaurant/bar for using the coupon is greater than 15 minutes)

toCoupon_GEQ25min:0, 1 (feature meaning: driving distance to the restaurant/bar for using the coupon is greater than 25 minutes)

direction_same:0, 1 (feature meaning: whether the restaurant/bar is in the same direction as your current destination)

direction_opp:1, 0 (feature meaning: whether the restaurant/bar is in the same direction as your current destination)

Y:1, 0 (whether the coupon is accepted)

Goals

Your marketing firm is trying to build general models to predict various consumer characteristics and behavior. Your clients are very interested in income, because they tend to market products to high-income individuals, and want to understand how income is related to various characteristics. Your goals will be 1a) to show whether income is significantly related to other variables in the dataset, 1b) to develop a quantitative prediction rule for income based on other variables, and 2) to predict whether coupons are accepted (Y=1 vs. Y=0) based on other variables.

For privacy reasons, variables such as age and income have been recorded in ranges or midpoint values. You need to make reasonable choices to deal with these annoying aspects of the data, which occur often in real life. For example, you might want to choose a midpoint of an income range to represent the income value, but that leaves a question of how to deal with "$100,000 or more." Also, many of the variables are string variables, but include quantitative information. Again, annoying but realistic.

Report:

The goal of the report is to investigate the questions above using the methods we've learned in this course.

You should write the report to an audience that has strong quantitative literacy skills but doesn't necessarily know a lot about statistics. The report should have the following sections:

- Introduction to the data: (basically, distill the information from the link, focusing on what you are going to investigate) You will want to explore the data first and identify your goals (see next step). Then come back to write the introduction.
- Goals: Here you should state what questions you are trying to investigate with your report. This should be detailed and, of course, centered around making inferences for our response variables in some way (perhaps discussing the average count of registered users on a weekday vs weekend, that kind of thing).
- Analysis:

  You may want to create new variables for use in your summaries and analyses.

  You should summarize and discuss all variables relevant to your goals and analyses done. These should be interesting summaries that show multivariate relationships (such as scatterplots, two-way contingency tables, logistic probability curves etc.) These summaries and plots should be embedded in the text, have nice descriptive labels/axes/etc., and generally be professional in appearance.

You should perform at least one contingency table analysis involving income, and a MLR analysis to predict income based on other variables. The MLR analysis of income must treat income as a quantitative variable.

You should report a multiple logistic regression analysis to predict Y from the other variables.

In regression modeling, you must perform at least some investigation of interactions, and at least one instance of stepwise model selection.

You should report at least three different kinds of hypothesis tests. These can overlap with other analysis, for example based on MLR output.

For each method used, you should describe the procedure briefly (including hypotheses being tested and test statistic as appropriate), state the assumptions needed for the interval/test to be valid (checking the assumptions where appropriate), and you should interpret all tests/predictions in the context of the study.

- Conclusions: Describe the overall conclusions you've made about your goals based on your data exploration and analyses. You should point to specific items from your analysis section that back up your conclusions.

SAS Programs:

You should be documenting all of the code you end up using with corresponding comments in a .sas program.

- You should create a permanent library (using code) in which to save your data.
- You should include all data manipulation steps in your programs.
- We should be able to recreate any of the graphs/summaries/analyses by running your code logically from start to end.
- You should include comments <u>after</u> your PROC steps used that give a very basic idea of what you are focusing on from that output/plot.

Rubric:

Item	Points	Notes
Intro	10	
Goals clearly defined and reasonable	5	
Data read in, manipulated, new variables created, etc.	20	Most annoying part!
Contingency table analysis correctly performed and described	5	Contingency table should include income as categorical/ordinal, and can use any other nominal or categorical variables of interest
MLR regression analysis	15	With income as a response
Stepwise analysis	5	This can be part of MLR
Hypothesis tests, assumptions tested, etc.	10	
Multiple logistic regression	10	Using Y as a response. If describing effects of variables, describe in terms of odds ratios.
Informative plots	10	
Conclusions clear and appropriate	10	
Total	**100**	

Notes on grading:

- For each item in the rubric, your grade will be lowered for each error (syntax, logical, or other) in the code or for each required description that is missing or lacking. The descriptions describe the coding part but also correspond to the relevant part of the write-up.

Final Note

Please have fun! These are real data that someone is interested in. If you can solve problems like these, you have learned the material in the course. I look forward to learning new things, and getting flashes of insight from your work.