Final Project

Project Goal:

The overall goal of the case is to provide actionable insight, based on the data available, as well as accurately predict which people (customers) will be expensive.

The dataset contains healthcare cost information from an HMO (Health Management Organization). Each row in the dataset represents a person.

Your team's goal is to understand the key drivers for why some people are more expensive (i.e., require more health care), as well as predict which people will be expensive (in terms of health care costs).

Hence, at a high level, you have two goals:

- Predict people who will spend a lot of money on health care next year (i.e., which people will have high healthcare costs).
- Provide actionable insight to the HMO, in terms of how to lower their total health care costs, by providing a specific recommendation on how to lower health care costs.

To do this, you will use all of the skills you have developed in the labs/homework to make sense of a novel dataset, to perform some essential analyses on the dataset, and to explain/document what you have done, and your insights generated.

Project Deliverables:

The analysis should include exploratory analysis (e.g., histograms, scatter plots), mapping visualizations and several machine learning techniques.

There are four deliverables for this project:

1. A report (e.g., word document) that describes the work done. This is a technical document and you should feel free to use technical terms. You must explain any visualization in your document.

2. The R code for your analysis (the team needs to figure out how to share code / work together)

3. A presentation of the actionable insight achieved. This actionable insight should be submitted in the form of a presentation, where the insights generated are explained. One specific recommendation must be given in the presentation. You will either present this in your lab, or you will record your presentation (e.g., a powerpoint presentation with your voice explaining the slides). Either way, it should be 10 minutes long. Your lab instructor will provide more details towards the end of the semester. Note: the client cares most about sensitivity, not overall accuracy.

4. A shiny app needs to be created and deployed on shinyapps.io – this app should read in a user selected datafile (i.e., new test dataset). This file has all the attributes that are available in your training dataset, except for the cost attribute. You will use a previously created and stored model to predict which people will be expensive. The app should also read in a second user selected file, which shows which people were actually expensive.

   Your goal is to have the best Sensitivity (a metric available in the confusion matrix).

   The shiny app should include:

   a. Some interactive exploratory analysis (based on the training dataset).

   b. Output of the Sensitivity (from the confusion matrix) for the new test dataset

   c. Output of the full confusion matrix of predictions

   → Note that the sensitivity and full confusion matrix should be generated with a previously generated, stored and retrieved predictive model (your best model), which is then used for the newly read in test data files.

The Data:

Here are the variables you will find in your data file:

- **X**: Integer, Unique identified for each person
- **age**: Integer, The age of the person (at the end of the year).
- **location**: Categorical, the name of the state (in the United States) where the person lived (at the end of the year)
- **location_type**: Categorical, a description of the environment where the person lived (urban or country).
- **exercise**: Categorical, “Not-Active” if the person did not exercise regularly during the year, “Active” if the person did exercise regularly during the year.
- **smoker**: Categorical, “yes” if the person smoked during the past year, “no” if the person didn’t smoke during the year.
- **bmi**: Integer, the body mass index of the person. The body mass index (BMI) is a measure that uses your height and weight to work out if your weight is healthy.
- **yearly_physical**: Categorical, “yes” if the person had a well visit (yearly physical) with their doctor during the year. “no” if the person did not have a well visit with their doctor.
- **Hypertension**: “0” if the person did not have hypertension.
- **gender**: Categorical, the gender of the person
- **education_level**: Categorical, the amount of college education ("No College Degree", "Bachelor", "Master", "PhD")
- **married**: Categorical, describing if the person is “Married” or “Not_Married”
- **num_children**: Integer, Number of children
- **cost**: Integer, the total cost of health care for that person, during the past year.

Data:

The data file is located at:

https://intro-datascience.s3.us-east-2.amazonaws.com/HMO_data.csv

Hints:

1. Make sure you define an attribute 'expensive', based on the cost attribute (that is the attribute the team should try to predict).

2. For exploratory work:
    a. Histograms and boxplots of numeric variables are typically useful.
    b. Producing tables of categorical response variables is often helpful.
2. Since there is geographic info, you should make some sort of map

3. Dividing the data in expensive and not-expensive subsets of people and using the grouping to visualize other attributes or build predictive models is often helpful.
    a. So, for example, Barplots of expensive (vs not expensive) people across different categories is often useful.

4. Remember to build models for predicting if a person is Expensive

5. For your recommendations (in the presentation):
    a. Provide a summary covering all of your results in language that is suitable for a manager to understand. Most managers do not know too much about statistics, so you probably should not quote terms like "R-squared" or "p-value" but rather describe your results in plain language.
    b. Your presentation should conclude with one substantive actionable recommendation to the managers.
    c. Important: Your recommendation MUST be connected with one or more of your data science results; it MUST NOT be based on your own personal experience with health, exercise, etc.

6. For your interactive shiny app:
    a. You should be predicting is someone will be expensive year (i.e., if that person will be expensive). You need to think about how you define 'expensive'.
    b. You need to make sure that you are reading a prebuilt model (not generating the model within the shiny app). It is just one model, so pick your best model.
    c. You need to be clear, in your output, what was the sensitivity from the new test data, as well as roughly, how well your team expected your model to perform (with respect to sensitivity).
    d. Example (sample) test datasets (one with the attributes, and one with the actual value of expensive for each person) will be provided. However, a much larger dataset will be used to test/evaluate your model. You will not get this larger test data set – it will be part of the evaluation (how your model works on the new data).