# SOA Predictive Analytics Exam
# Sample Problem

## Background

Life insurance companies continually seek new ways to deliver products to the market. Those involved in product development wish to know "who buys insurance and how much do they buy?" Analysts can readily get information on characteristics of current customers through company databases. To expand market share, the focus is on potential customers, those that do not necessarily have insurance with the company.

You are an actuary in the product development area of PA Insurance (PAI). Your department has obtained external data from the Survey of Consumer Finances, a nationally representative sample that contains extensive information on assets, liabilities, income, and demographic characteristics of those sampled. The sample that has been obtain contains data on 500 households with positive incomes that were interviewed in the 2004 survey.

Your supervisor, who is an actuary with significant analytics knowledge, has determined that there are two variables of interest to be used as the target variable. One is TERM_FLAG. This variable is 1 if the individual owns a term life insurance policy and 0 otherwise. The second variable is FACE, which is the amount of the death benefit for the term life insurance policy. For respondents without term life insurance, this value is set as 0. The appendix provides a description of the other variables in the dataset that have been identified as potential predictors.

You have been asked to construct models that can be used to provide insights regarding who will buy this type of insurance and, for those who buy it, how much are they likely to purchase. Your report should accomplish two things:

1. Provide evidence to your supervisor that you have constructed appropriate models given the available data.
2. Provide insights to the marketing department regarding drivers of purchasing behavior.

## Deliverables

### Report

Your report should be a single Word document with the following sections as listed. Due to the hybrid audience, sections 2-4 should be written with your supervisor as the audience while sections 1 and 5 should be written with the marketing department as the audience.

1. Executive Summary: The audience is an executive in your marketing department who wants to know what you have learned from your model-building exercise.
2. Data Exploration, Preparation, and Cleaning: Present tables and graphs that will help your supervisor understand the nature of the variables and how they relate to the target variable. Also, ensure that the data is ready for use in building models.
3. Feature selection: Consider transformations, interactions, and other approaches to obtain a set of features to feed into the selected models. While you may learn more during the model-building process, it will be sufficient to work with the set of features determined at this step.[1]
4. Model selection and validation: Keep in mind the business problem and that the results will need to be communicated to the marketing department. Two models will be constructed:
    a. A decision tree, with extensions as appropriate, to predict whether an individual owns a term life insurance policy; and
    b. A generalized linear model, with extensions as appropriate, to predict the amount of the death benefit if a term life insurance policy is owned.

    In constructing the models, choose an appropriate validation method and determine which combination of features provides the best outcome. Explain your decision-making with respect to model construction and validation. For each of the two models indicate an alternative type of model that may have been used and whether or not it would be a better option based on the particular problem you are solving. Do not fit alternative types of models.
5. Model explanation: Describe the model and how the selected variables relate to the target variable. This explanation should be for the marketing department. Any caveats should be communicated at this time.

Note that there is no conclusion section. While this is commonplace, for this project there is no need to spend time repeating information that has already been communicated.

Appendices can be used, but are not required. Generally, items that are important, but would disrupt the narrative, can be moved to the appendix. However, it is important to let the reader know as they read the report what is in the appendix so they can decide if it is worth reading.

## Code

You are to also deliver your code in the form of an Rmd file. Some guidelines are:

- Assume that the graders will run your code. At a minimum it should work. Note that if as part of your work you save and then reload data files, be sure to either ensure the code has created and saved them or the files are separately uploaded.
- Comment your code. Help the graders (and your hypothetical supervisor) understand what your code is trying to accomplish.
- Do not reference the code in your report. The report should be self-contained. However, the Rmd file can reference the report.

[1] Projects for the PA exam must be able to be completed within the five-hour time period. Here is a case where you are provided instructions that reduce the work load versus a thorough analysis. In particular, the iterative nature of building models will need to be limited.

## Additional Files

You can upload any other files that you believe will support your report. Be sure that there is an explanation (other than data files as noted above) in your report regarding why readers should look at these files. There is no expectation that additional files will be needed, but the opportunity is available.

## Appendix: Data Background and Data Dictionary

The dataset comes from the repository established by Professor E. W. Frees to augment his textbook *Regression Modeling with Actuarial and Financial Applications*. The datasets can be found at this web page: http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html.

Not all of the variables in the set are being used here and the TERM_FLAG variable has been derived from the FACE variable. It is important to note that this is part of a much larger survey. Many of the details are about the person who completed the survey, which may not be the person who is insured by the term life policy. The data dictionary follows:

| Variable Name | Description |
|---|---|
| FACE | Amount of the death benefit |
| TERM_FLAG | 0 if no term insurance purchased, 1 if term insurance purchased |
| GENDER | 0 = Female, 1 = Male |
| AGE | Age in years of the respondent |
| MARSTAT | Marital status of the respondent: 1 = married, 2 = not married, but living with a partner, 0 = all other |
| EDUCATION | Number of years of education of the respondent |
| SAGE | Age of spouse/partner |
| SEDUCATION | Number of years of education of the spouse/partner |
| NUMHH | Number of household members |
| INCOME | Annual household income |
| CHARITY | Charitable contributions |

Note: A value of zero could indicate that the response is 0 or could indicate a missing or inapplicable value.