Rory Creedon

Predicting Risk Class

Using life insurance application data to predict mortality risk class

The Problem: Buying Insurance is Hard

The Customer

- There is a trillion dollar life insurance gap in the USA
- Partly this is because buying insurance is so hard:
    - Application process requires “fluids” as well as survey responses
    - Extremely time consuming
    - Invasive
    - Not in line with current purchase journeys (e.g. online, rapid, quick delivery)
- Customer purchase experience could be significantly improved if applications were based on predictive models of a parsimonious set of data points.

The Insurer

- Insurers categorize people into risk classes based upon their expected mortality (how long they will live).
- The risk class tells the insurer how much to charge in premium.
- Currently most life insurers rely on an extensive application questionnaire that covers lifestyle and health of applicants.
- This is often supplemented by physical examinations.
- The process is very costly for the insurer (up to $200).
- The insurer could save money and attract more business (passing on cost savings, improving customer experience) if risk class could be predicted using a small set of data points.

The Data: Prudential Application Data

The Dataset

- In 2015 the Prudential life insurance company released a dataset to the Kaggle community as part of a competition to build a model that predicts risk classes based on a limited set of insurance application data.
- The response variable is an ordinal variable (0-8) which represents the risk classes. Typically insurers measure risk relative to the 'standard life'. There are risks that are better than standard (e.g. super preferred) and risks that are worse than standard (e.g. tobacco).
- The predictors are applicant attributes that have been drawn from a self-reported questionnaire that asks information about applicant health and lifestyle.
- The challenge is to use the predictors to predict the values of the response variables.
- The winning entry did this with a 0.679 degree of accuracy.
- The data is divided into a training set and a testing set.

Data Fields

- There are over a hundred data fields available, and they can be broadly characterized as follows:
  - Basic applicant physical data (height, weight)
  - Employment history
  - Insurance history
  - Medical history
  - Family history
- The majority of the variables are categorical. Height, weight and related fields are continuous.

Initial Hypotheses & Defining Success

Hypotheses

- Socio-economic variables are consistently found to correlate with mortality. These variables are proxies for behaviors that are not generally recorded on applications but influence mortality. Therefore we expect employment related variables to be good predictors. Of interest could be salary, employment duration, frequency of job changes, continuous employment.

- Medical variables can reveal health issues and proxy for lifestyle choices that influence mortality (and therefore risk class). In particular diabetes, cancers, and HIV are typically indicative of higher mortality risk.

- Calculated fields relating to physical measurements are generally predictors of mortality, e.g. BMI etc. These can be computed from the data.

- I have access to underwriters at my work and I believe I can ask them to assist me in identifying features that I can pull from this data.

Defining Success

- The winning entry on Kaggle predicted risk class with a 0.679 degree of accuracy. If I can approach or surpass this level of precision then the exercise will be a success.

- More broadly understanding how to think about these data and how to model them will allow me to take the learnings back to my job where this is a live issue that is being worked out throughout the organization.

- I will present my findings to my colleagues as well as my classmates