

DNA HACKATHON

Predictive Underwriting Systems (Insurance Industry)

SPOC: Anand Nath Jha, 1039220, anandnath.jha@infosys.com, 9741258040

Context

Underwriting (UW) is perhaps one of the most critical functions of Insurance Industry. Significant amount of cost and efforts are dedicated for this. With the advent of widespread use of Artificial Intelligence, there has been lot of interest in leveraging AI for UW. It has the potential to become a potent aid for Underwriters - less need of UW Resources per application, expedited UW processes, significant Cost and time savings.

Problem Description

Prudential Life Insurance made available their UW dataset for 60,000 customers. The dataset had 8 risk classes, coded as 1 to 8, with 1 being the highest risk. There are 126 inputs available, pertaining to health, weight, medical test, employment history, family history etc. The variables are masked by naming them in generic fashion like MEDICAL\_KEYWORD, EMPLOYEMENT\_INFO etc. The objective is to develop a predictive model which can classify these risk classes with very high degree of accuracy, maintaining reasonably high precision and recall within each classes.

Challenges

* Data made public by Prudential Life Insurance
  + ~ 60,000 records
  + 8 Risk Categories (Coded as 1, 2, 3, …, 8)
* 126 Variables
  + Height, weight, BMI etc.
  + Family History (4)
  + Employment Information
  + Insurance History
  + Product Information
  + Medical History (41)
  + Medical keywords (48)
* Highly imbalanced data set for Risk Categories as well as Input variables
* Risk classes 4 & 6 constitute merely 2.4% and 1.7% of total records respectively.
* E.g. MEDICAL\_KEYWORD\_48 has 5% as code 1 and rest as 0
* Same observed for multiple other variables

Data

There are 126 input variables, with 8 risk classes. The column headers do not provide explicit names of these variables and coded as EMPLOYEMENT\_INFO\_1, EMPLOYEMENT\_INFO\_2 etc. The data set is attached here.

Approach

* The analysis was performed in R.
* H2O in R was used with balance\_class = TRUE option
* H2O also makes the computation faster.
* After few preliminary analysis, Random Forest model was used.
* Random Forest was used in a two stage manner.
  + First iteration was used for variable screening (most important differentiators for risk classes).
  + The list of important variables from 1 above were used finally for Classification.
* The results are provided below.

