Data Science in Action Proposal

Zhiyu Quan, University of Connecticut[*]

22 February, 2018

1 Introduction

1.1 Description and Background

Fourth Actuarial Pricing Game as part of a research project conducted by Arthur Charpentier. From February 2018 until May 2018, the competition provides real historical(two year) contracts and claims data(about 10K contracts) to train models and offer premiums.

1.2 Problem Statement

The goal is to offer a premium for all records in the pricing dataset with the model information. There will be limits on the complexity of first submission while for final submission there are no restrictions.

1.3 Time Line

- Register by February 28th 2018
- Receive a training dataset by February 28th 2018
- First submission by April 9th, 2018
- Final submission by May 14th, 2018

2 Model Candidates

Insurance is risk sharing in simple. We define a single claim severity as Y_i, $i = 1, \ldots, N$, here N is number of claims. Finally we can define insurance premium as π,

$$\pi = \mathbb{E}_{\mathbb{P}}(\sum_{i=1}^{N} Y_i) \quad \text{here } \mathbb{P} \text{ is risk space}$$

Ideally, with insurance price differentiation, we have

$$\pi(r) = \mathbb{E}_{\mathbb{P}}(\sum_{i=1}^{N} Y_i | \mathbb{R} = r) \quad \text{for certain risk } \mathbb{R}$$

In reality, with given data, we have risk explanatory-variables $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$, so

$$\begin{aligned} \pi(\mathbf{x}) &= \mathbb{E}_{\mathbb{P}}(\sum_{i=1}^{N} Y_i | \mathbf{X} = \mathbf{x}) \\ &= \mathbb{E}_{\mathbb{P}}(N | \mathbf{X} = \mathbf{x}) \mathbb{E}_{\mathbb{P}}(Y_i | \mathbf{X} = \mathbf{x}) \quad \text{assume independence between } N \text{ and } Y_i \end{aligned}$$

[*]PhD Candidate, Department of Mathematics, University of Connecticut, 341 Mansfield Road, U-1009, Storrs, CT, USA, 06269. zhiyu.quan@uconn.edu.

We build predictive models to estimate $\mathbb{E}_{\mathbb{P}}(N|\mathbf{X}=\mathbf{x})$ and $\mathbb{E}_{\mathbb{P}}(Y_i|\mathbf{X}=\mathbf{x})$ from given data. Under independence assumption, traditional statistics methods approximate π using $\pi(\mathbf{x})$ by the law of large number theorem. While, many empirical studies show that the independence assumption is not valid which leading to a biased premium estimation. I will explore some nonparametric techniques to find a remedy for this situation. Here are some model candidates:

- Traditional
  - Tweedie GLM
  - Two-stage model
  - Elastic net
- Tree-base model
  - Regression tree
  - Conditional inference tree
  - Random Forest
  - GBM
- GAM
  - GAMboost
- Neural Network
  - LSTM
- Clustering
  - Segmentation on training data

3 Data

The **training dataset** will be based on two years of data on household policies. There will be information about the insurance policy from underwriters, as well as, claims data. The **pricing dataset** with only underwriters information and no claims.

4 Future Work

- EDA
- Data Cleaning and Feature Engineering
- Modeling Tuning
- Prediction Accuracy

References