Project Proposal

Background and motivation: A core piece of the insurance business is risk assessment, selection and segmentation. This ensures that each risk is adequately priced. In other words, the premiums paid by each insured are proportional to the corresponding insured risk characteristics and insurance coverage provided. Inadequately pricing insurance policies may lead to very severe consequences ranging from poor profitability to insolvency. In fact, according to a 2009 report by the Property and Casualty Insurance Compensation, from 2000 to before 2009, a total of 115 insurance companies, in Canada and the USA, became insolvent and exited the market due to poor pricing.

Generalized Linear Models (GLMs) are widely used in the property and casualty insurance industry for risk assessment, selection and segmentation. However, for life insurance policies selection and classification, Survival Analysis Models such as Cox Proportional Hazard (CoxPH) models are usually employed. The biggest strength of these two modeling techniques is their ability to yield transparent, explainable and interpretable models. In the context of insurance risk segmentation and pricing, transparency and interpretability are usually key regulatory requirements. However, the model transparency and interpretability of GLMs and CoxPH also constitute their weakness because they come at the cost of certain rigid or sometimes unrealistic model assumptions that may result to oversimplification of the relationship between risk characteristics and the corresponding risk level presented by the insurance applicant. This in turn can lead to sub-optimal risk selection, classification and segmentation decisions.

Modeling techniques based on ensemble of regression trees such as Random Forest or Gradient Boosting Machines (GBM) have proven very effective is capturing complex relationships in data. The same is true for modeling methodologies based on neural networks with sufficiently high number of hidden layers, otherwise known as deep learning. However one of the biggest weaknesses of modeling techniques such as GBM or deep learning , in the context of insurance risk modeling, is their lack of guarantee to produce a model that is fully transparent and interpretable.

Goal/Objectives: The objectives of the project are:

- Investigate ways to combine the two categories of modeling techniques mentioned above in a way that yields highly predictive models that can capture complex relationships in the data without sacrificing transparency and interpretability
- Develop a piece of software (e.g. R/Python library) that implements the proposed methodology
- Perform simulation studies to demonstrate that the software/implementation of the proposed methodology works as intended
- Use the developed software to solve the 2014 NYS Behavioral Risk Factor Surveillance System (BRFSS) data Challenge

Expected Learning Outcomes: At the end of the project, students are expected to:

- Gain a better understanding of statistical and machine learning theory
- Improve their programming skills in R or Python
- Develop model building and communication skills
- Develop understanding of the application of machine learning and statistics to insurance risk modeling

Potential Project Outcome:

- Technical report or journal article
- R or Python based software

2014 NYS Behavioral Risk Factor Surveillance System (BRFSS) data Challenge:

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual statewide telephone surveillance system designed and funded by the Centers for Disease Control and Prevention (CDC), and conducted by the NYSDOH Division of Chronic Disease and Prevention, Bureau of Chronic Disease Evaluation and Research. The BRFSS collects data on preventive health practices and risk behaviors that affect chronic diseases, injuries, and preventable infectious diseases. Examples include tobacco use, health care coverage, HIV/AIDS knowledge and prevention, physical activity, and consumption of fruits and vegetables. Demographic information is also collected to permit analyses of specific populations. While all data collected are self-reported, some variables are calculated based on given responses. For example, obesity is calculated based on the respondent's reported height and weight. Current smoking status and leisure time physical activity are also calculated variables.

Overall health and pre-existing medical conditions are part of the factors that insurance companies evaluate in order to determine your insurability and related cost of insurance coverage. The current industry practice of assessing health risks is to model each individual health outcome in isolation of the others. For instance, researchers may model risk of obesity in isolation of diabetes risks. However it is often the case that these two health risks are not independent of one another.

The purpose of this challenge is to develop a model/algorithm that:

- allows modeling of the following health risks: DIABETES(DIABETE3), OVERWEIGHT OR OBESE CALCULATED VARIABLE(RFBMI5), COMPUTED SMOKING STATUS(SMOKER3) in a way that leverages potential dependencies among these three health outcomes.
- determines early warning signs of diabetes and identify top behavioral and demographics risk factors associated with diabetes. In addition, provide a qualitative and quantitative description of the relationship between each of the top risk factors and risk of diabetes.
- determines top behavioral and demographics risk factors associated with obesity. In addition, provide a qualitative and quantitative description of the relationship between each of the top risk factors and risk of obesity.
- identifies respondents who answered that they do not smoke while they actually do.
- identifies respondents who answered that they do not suffer from diabetes when they actually do.
- provides deep insights on the relationship between diabetes, obesity and smoking status.