Kaggle 2014 – My Experience

Team name: LM_Lucinda - Team members: Marcos Aguilera Keyser - Country Office: Liberty Spain

1st/21 LMG teams and 36th/634 teams overall position – top 10%

Introduction

The **Liberty Mutual Group – Fire Peril Loss Cost Kaggle competition** that took place last summer was my first participation in Kaggle.

The algorithms more commonly used in a Kaggle competitions are **non-parametric** because are more flexible and therefore with greater prediction accuracy. **Random forests** is one of the most popular. **R software** is also the most popular technology in Kaggle contents.

I am feel more comfortable working within the **GLM framework** using **SAS software**. So, I decided to use a **parametric algorithm** such as a **Generalized Linear Mixed Model (GLMM)**. Also, I used SAS STAT and SAS Enterprise Miner as my technology.

The Problem

The problem

Within the business insurance industry, **fire losses** account for a significant portion of total property losses. **High severity** and **low frequency**, fire losses are inherently volatile, which makes modeling them difficult. In this challenge, your task is to predict the target, a transformed **ratio of loss to total insured value**, using the provided information. This will enable more accurate identification of each policyholder's risk exposure and the ability to tailor the insurance coverage for their specific operation.

Challenges

1. **Model specification**: which model use to explain very low number of losses 1188 in 452061 that is around 0.00263 of losses in the training data.
2. **Variable selection**: how select a parsimony model with more than 300 variables?
3. **Data imputation**: only a 34% are complete cases in the training data. How to handle more than 300 variables with missing data?

Objectives

- **Prediction accuracy** is the main objective of the competition
- **Avoid over-fitting** is the most critical issue to overcome during the competition

Methods

Algorithm

Generalized Linear Mixed Model (GLMM) with a **cross-classified data structure** with two **random effects.** The error function was a **Tweedie**. I used the **GLIMMIX procedure** in SAS

Variable Binning and Missing Imputation

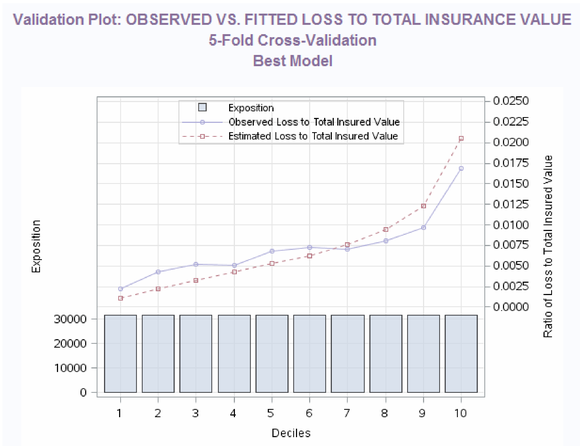
I used **decision trees** as an auxiliary tool in order to perform variable binning and some missing imputation. **SAS Enterprise Miner** was a good tool for these tasks.

Variable Selection

With around 300 hundred variables, the variable selection step was critical. I used different **stepwise algorithms**. At that moment I had never heard about **Elastic Nets** as a successful variable selection method. After Kaggle I started to use Elastic Nets and now I am a fun.

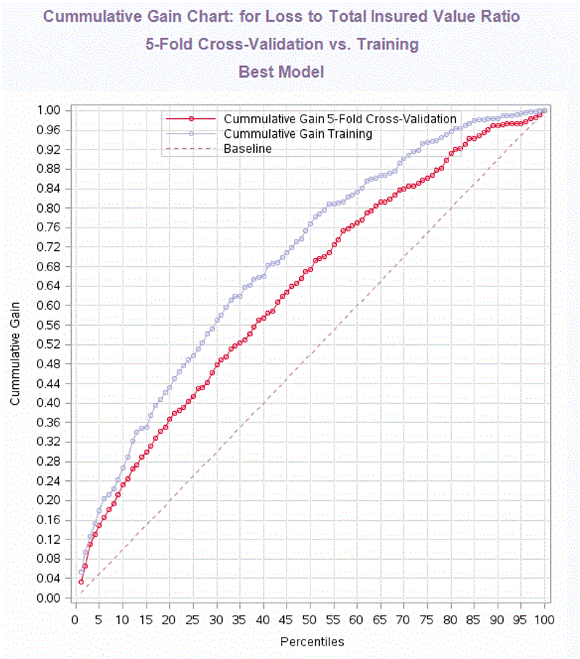
Model Validation

The most challenger issue in a Kaggle competition is to avoid **over-fitting**. As the great majority of the competitors I used **five-fold cross-validation** in order to avoid over-fitting.

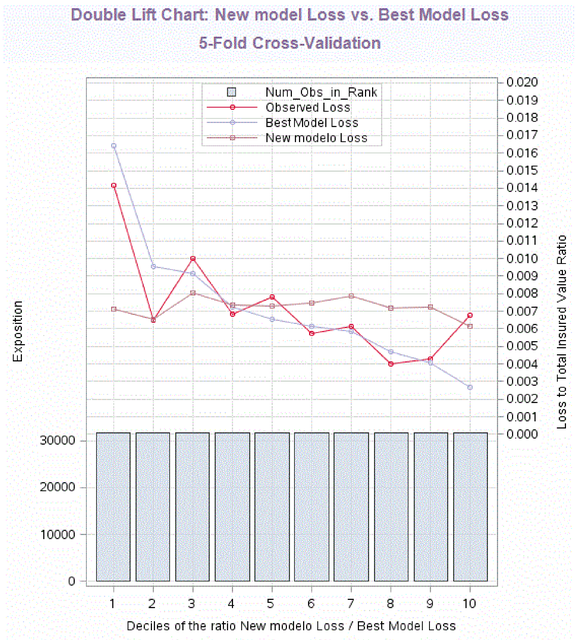


The above image measures how well the model fits the observed data using cross-validation.

The next image shows the **Gini index** for the cumulative gain curve on the training data is **38.52% vs. 26.49%** under five-fold cross-validation. The Gini index under cross-validation was critical in order to choose my best model and try to avoid over-fitting.



The next chart shows how critical was the inclusion of **two random effects as intercepts**. The "New model" represents a GLM without mixed effects and the "Best model" is the GLMM with mixed effects.



Lessons

- A **parametric algorithm** is not too far from the best possible algorithm – the winner of the public contest. The GLM framework used in ratemaking in the insurance industry works quite well
- The use of GLMM in order to deal with spare data and **lack of credibility** definitely was critical
- A **careful reading of the problem description** was very important. The problem description mentioned clearly that one of the variables was in **hierarchical structure**. It was a clue in order to use mixed models
- **A pragmatic approach** was important too. I decided not to waist time trying to learn new algorithms such as Elastic Nets from scratch, instead, I used those algorithms with which I was already familiar

Conclusions

- Competing in my first Kaggle competition during de summer of 2014 has been **a fantastic experience**.
- I had the chance to face **one of the most complex and interesting problems** in my professional career.
- The Kaggle competition has **become in my best way to learn** new thinks and polish my current skills in predictive modeling
- Kaggle gave me for the very first time the opportunity to compare my skills and experience as a predictive modeler with other data scientists

References

Klinker, Fred (2010), *Generalized Linear Mixed Models for Ratemaking: A Means of Introducing Credibility into a Generalized Linear Model Setting*. Casualty Actuarial Society E-Forum, Winter 2011-Volume 2

Breiman, Leo (2001), *Statistical Modeling: The Two Cultures.* Statistical Science, Vol. 16, No. 3 (Aug., 2001), 199-215

SAS Institute training courses:

- *Mixed Models Analyses Using SAS*
- *Statistical Analysis with the GLIMMIX Procedure*
- *Multilevel Modeling of Hierarchical and Longitudinal Data Using SAS*