# Enhancing the Generalized Linear Modeling Approach with Machine Learning Technique

Jie Dai, FCAS, CSPA

**Abstract**

With the development of the machine learning (ML) technique and broad successful application, machine learning is becoming more and more popular for data analytics in many industries. Insurance is no exception, and machine learning techniques are used to build predictive models in Claims (Fraud, subrogation models), Marketing (Segmentation, cross sell model, recommendation models), and Underwriting. However, for pricing models, Generalized Linear Models (GLM) still dominates given its easy interpretation and well-established frame work. Using a machine learning method to enhance the GLMs model is a challenge to the insurance industry especially for actuarial modeling. This paper will discuss some potential ways to enhance the GLMs model with tree based machine learning techniques and give a case study on territorial analysis, which would show significant improvement on the predictive nature of the GLM model.

**Keywords**. Machine learning; territorial analysis; generalized linear modeling.

## 1. INTRODUCTION

Machine Learning is a sub-set of artificial intelligence where computer algorithms are used to autonomously learn from data. Machine learning (ML) is getting more and more attention and is becoming increasingly popular in many other industries. Within the insurance industry, there is more application of ML regarding the claims and underwriting disciplines. There is little in actuarial literature on ML, and none is in pricing modeling.

In the early 1970s, Nelder and Wedderburn coined the term generalized linear models (GLM) for an entire class of statistical learning methods that include both linear and logistic regression as special cases. In the last two decades, GLMs have been widely in use in P&C insurance to classify risks and determine rate structures. However, standard GLMs do have several shortcomings, most notably [1]:

- Predictions must be based on a linear function of the predictors;
- GLMs exhibit instability in the face of thin data or highly correlated predictors;
- Full credibility is given to the data for each coefficient, with no regard to the thinness on which it is based;
- GLMs assume the random component of the outcome is uncorrelated among risks;
- The exponential family parameter $\emptyset$ must be held constant across risks;
- GLMs only can identify simple and global interactions, which are the interactions between

all levels of two predictors. For identifying complex interactions with GLMs, the manual process would be non-trivial.

Also, another challenge of using GLMs includes the selection of predictors from large volume of variables candidates.

In mid 1980s Breiman, Friedman, Olshen and Stone introduced classification and regression trees, which for the first time made fitting non-linear relationships computationally feasible. Since then there are more algorithms (like neural nets, random forests or gradient boosting) that have been developed and widely used in other industries or disciplines [2]. Those methods don't have the shortcomings noted above, and therefore able to produce strong models that have the potential to yield more accurate predictions. However, using those methods directly would entail a huge loss of interpretability, which is critical for many actuarial applications.

This paper will present the ways to enhance the GLMs with ML technique in variable selection and feature engineering. In addition, we will look at an application in sewer backup modeling that shows significant improvement of the model results with the new features created through ML. However, for reasons of confidentiality, we are not able to share detailed data and quantitative results in this paper.

### 1.1 Research Context

With more and more data being available for pricing models, the challenge arises to reduce the number of predictors to improve the prediction accuracy and interpretability. Stepwise selection (forward, backward and/or hybrid) are widely used in GLM modeling practice. Recently, shrinkage methods like Lasso (least absolute shrinkage and selection operator) have become more popular because it can be a more efficient method that produces more interpretable models that involve only a subset of the predictors. The third method to reduce variables or dimensions is to create predictors from the original raw predictors. Principal Components Analysis (PCA) is the most popular approach in deriving a low-dimensional set of features from a large set of variables. Insurance score is a major rating variable introduced to personal lines insurance [3] in the late 1980's and 1990's. This variable is derived from dozens of selected/created credit variables (from initially thousands raw variables) to predict insurance loss risk by using linear regression and/or ML. Another popular rating variable in auto insurance which is created from dozens of raw vehicle characteristic variables is auto symbol. Both variables are easier to interpret and reduce the dimension significantly compared to using the raw underlying variables.

Interaction identification is a challenge in GLMs modeling in practice, especially for the interaction

among more than 3 variables. Some ML techniques naturally will include all the possible interactions between variables. Creating new features based on the ML techniques to replace the underlying raw variables would not only reduce the number of variables but also significantly improve the predictive power of the GLMs model.

To do the territorial analysis for a sewer backup modeling, 14 geographic variables are studied which are not predictive in the model. A score variable was created from these 14 variables, and a territorial definition was created from census block group with the help of the 14 geographic variables. The score variable can be used in underwriting and pricing. Both new features would improve the predictiveness of the GLM model significantly.

### 1.2 Objective

Our objective is to use the feature created with ML from some underlying variables to improve the predictive power of the GLMs. Those new features should be like vehicle symbol or credit score which can be interpreted to a certain degree.

### 1.3 Outline

The reminder of the paper proceeds as follows. In Section (2.1), we discuss the sewer backup data and modeling. In section (2.2), we discuss the territorial analysis, and especially the challenge for sewer backup loss data. In section (2.3), we discuss the tree based supervised learning methods in ML. In section (2.4) we introduce the double lift curve for the model comparison. In section (3.1) we present the result that shows even if the raw variables are not good predictors, the score produced from them through ML can be very predictive. Finally, in section (3.2) we present the model comparison with and without the boundary, which shows the significant improvement with the boundary variable. The boundary variable is created by grouping census block group.

## 2. BACKGROUND AND METHODS

### 2.1 Sewer Backup Modeling

The sewer backup loss modeling dataset included observations with sewer backup coverage endorsement. Since this loss is highly correlated with location, the territorial analysis should be important. To do the territorial analysis and create the boundary, we tested 14 geographic variables from US census data. The 14 geographic variables include Water Surface Elevation, Average Travel Time, Average Household Size, Average Number of Vehicle, Population Growth in 5 years, Average Age etc. Unfortunately, none of them showed predictive power. It also is difficult to create a territorial

boundary with a spatial smoothing method. For this study, we only present the result for frequency models.

### 2.2 Territorial Ratemaking and Boundary

For territorial ratemaking, the first phase is to establish territorial boundaries [4]. Census block group (CBG) is selected as the basic geographic unit due to its small size and relative stasis over time. The current approach to create the boundaries include the following steps [4]:

- Create geographic estimator on CBG with geographic indictors by building a GLM model using a variety of non-geographic and geographic explanatory variables;
- Applies spatial smoothing techniques to the geographic residuals to see if there are any patterns in the residuals and those residuals can be used to adjust the geographic estimators to improve overall predictive power of the model.
- Once the geographic estimators are calculated for each CBG, the CBG can be grouped into territories.

Our proposed approach is to create the CBG estimator by building a GBM (gradient boosting machine) model on the residual of the GLM model by using the 14 geographic variables. The GLM model is created using non-geographic and geographic explanatory variables. With the help of smooth weight of evidence (SWOE) [5] we transferred the categorical variables (CBG) into an interval variable, and then created a boundary based on the decision tree, we grouped the census block group into 19 levels.

### 2.3 Tree-Based ML Techniques

Tree based methods partition the feature space into a set of rectangles, and fit a simple model (like a constant) in each one [6]. Assume our data consists of p inputs and a response, for each of N observations: $(x_{i,}\ y_{i,})$ for i=1,2,…N, with $x_i = (x_{i1,}x_{i2,\ldots,}x_{ip,})$. For regression tree, if we have a partition into M regions $R_1, R_2 \ldots R_M$, and we model the response as a constant $C_m$ in each region:

$$f(x) = \sum_{m=1}^{M} c_m\, I(x \epsilon R_m\,) \tag{2.1}$$

It is easy to see that the best $\hat{c}_m$ is just the average of $\mathrm{y_i}$ in region $R_m$ :

$$\hat{c}_m = ave(y_{i,}|x \epsilon R_m). \tag{2.2}$$

The big advantage of a tree based ML technique is that it is easy to interpret, and easy to implement. It is still a great tool for identifying interaction or as a supplement analytic tool for other more advanced techniques. One major problem with trees is their high variance [6]. A small change in the data can result in a very different series of splits, making model chosen somewhat precarious. To reduce this variance, several tree based algorithms have been developed, which are more predictive and would reduce the possibility of over fitting the model. Among them, the two most common of these techniques used are boosting and bagging.

A Gradient Boosting Machine (GBM) is a generalization of tree boosting that attempts to mitigate some problems with other boosting methods like speed, robustness and interpretability [6]. The generic algorithm for the GBM is listed here [6]:

$Initialize\ f_{0(x)} = argmin_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)$

For m=1 to M:

For i=1,2,…N compute $r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$

Fit a regression tree to the targets $r_{im}$

For i=1,2,… $J_m$ compute $\gamma_{jm} = argmin_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$

$$f_m(x) = f_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}) \qquad (2.3)$$

Where $L(y_i, \gamma)$ is the loss function, and the parameter $\nu$ can be regarded as controlling the learning rate of the boosting procedure. Both $\nu$ and M control prediction risk on the training dataset. Smaller values of $\nu$ lead to larger values of M for the same training dataset, so that there is a tradeoff between them. When M is large, the computation becomes expensive and would take a long time to run. To our experience, $\nu$ may vary from 0.01 to 0.15 and M can be from 50 to hundreds depending on the dataset. We run the model with SAS enterprise miner, other tools or package (R or Python) of gradient boosting may choose different $\nu$ and M to get the best result.

Random forest is a substantial modification of bagging that builds a large collection of de-correlated trees, and then averages them. On many problems, the performance of random forests is very like boosting, and they are simpler to train and tune, and random forest is easier to parallelize and robust to overfitting. That's why random forest is also popular in ML application. However, in the author's experience, we do see GBM outperform random forest in many insurance applications if it is well tuned.

The advantages of tree based models over GLM include but are not limited to:

- No assumption of model structure which would be learnt from data;
- Easy implementation of complex and/or multiple way interactions;
- Easy to deal with missing values;
- Built-in feature selection;

### 2.4 Double Lift Curve

For modeling comparison, a double lift curve is a simple method to directly compare the predictive accuracy of two models. Here we use EMBLEM's model comparison function to compare two model's performances. The X axis is the bucketed ratio of indications of the two models, and the graphs will show the two models' average indications in those buckets and the average of actual observations in those buckets. The "winning" model would be the one that matches better the observed frequency in each bucket. In all the following models, we split the dataset into 80% and 20% randomly as training and validation dataset, and the double lift curves are created on the validation dataset using EMBLEM.

## 3. RESULTS AND DISCUSSION

### 3.1 Geographic Variable Score

To show the idea that the complicated interactions are important and are missed sometimes in the GLM modeling, we built two Frequency models: model1 is the model with all the current rating variables plus 14 geographic variables; model2 is the model with all the current rating variables plus a geographic variable score (geoonly14), which is created based on the 14 geographic variables with GBM.

Fig 3.1 shows the double lift curves for model 1 and 2. Based on these results, we see that model 2 is significantly better in predictive accuracy. This result shows a case where even when the individual variables are NOT predictive; the combination of the variables can be very predictive because of the complicated interactions between those underlying variables. Looking for interactions among 14

variables with many levels would be a non-trivial work and especially difficult because we generally have no prior knowledge regarding the potential interactions between geographic variables. And we are not able to identify/include interactions among 3 or more geographic variables in GLM with EMBLEM.

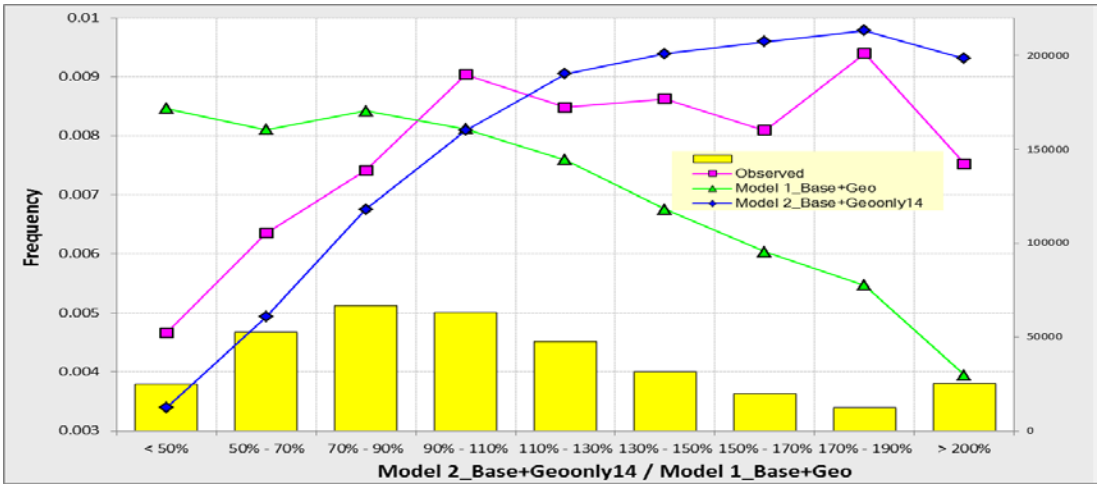


Fig 3.1 Double Lift Curve for the Model Comparison for Geo Variables and Score

### 3.2 Sewer Backup Territorial Boundary

For territorial analysis, the current method is to use geographic variables to create the indication for the census block group (CBG) with GLM modeling, and then use the classifier of EMBLEM to do the spatial smoothing and correction (if there is pattern in the residuals). However, it is very difficult to find the pattern in the residuals, and thus the correction is also subjective in practice. In theory, we can use CBG as the variable to create the indication for territorial rating. The hurdle would be how to group the more than ten thousand levels of CBG. We use SWOE to recode the CBG and with the help of a decision tree model on the GBM model output, we can create the CBG group which could be used in the territorial rating directly. We produced the 19 CBG groups and incorporated it into the base model for our sewer backup classification GLM model. Fig 3.2 shows the comparison of the two models: Model 1: Base model (current rating plan) plus the geographic variables; Model 2: Base model plus the 19 CBG grouping variable (Territorial Boundary). Model 2 shows significant improvement over Model 1 in predictive accuracy.

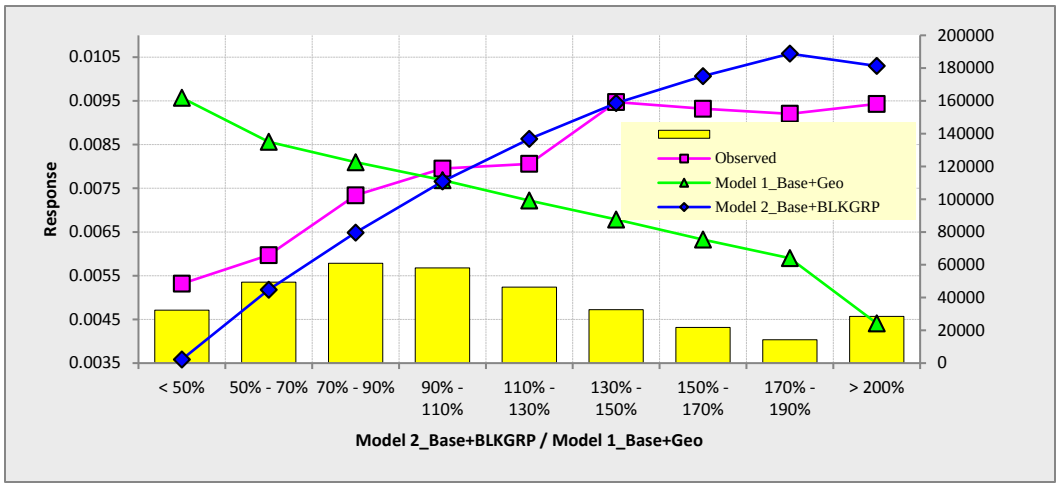


Fig 3.2 Double Lift Curve for the Model Comparison for Geo Variables and Territorial Boundary

## 4. CONCLUSIONS

With the development of the advanced modeling techniques, there are more and more data and variables available for pricing. It is a challenge to select variables and/or extract information from those raw variables to build a model which is more accurate in predictive power and still interpretable. To keep the GLM framework intact, the methods presented in this paper show the potential ways to incorporate advanced analytical techniques, especially machine learning, into the variable selection and dimension reduction procedure, which may significantly increase the predictive power of the model. This method can be applied to develop vehicle symbol, territorial boundary and other risk score variables.

## 5. REFERENCES

[1] Mark Goldburd, Annand Khare and Dan Tevet, "Generalized Linear Models for Insurance Rating", CAS Monograph Series Number 5.
[2] Gareth. James, Daniela Witten, Trevor J. Hastie and Robert John Tibshirani., An Introduction to Statistical Learning, Springer Science & Business Media, 2013.
[3] Conning Report, "Insurance Scoring in Personal Automobile Insurance – Breaking the Silence", Conning Report, Conning, (2001).
[4] Geoff Werner, Claudine Modlin, Basic Ratemaking, Fifth Edition, May 2016.
[5] Satish Garla, Goutam Chakraborty, Andrew Cathie, "Extension Nodes to the Rescue of the Curse of Dimensionality via Weight of Evidence (WOE) Recoding", SAS Global Forum 2013.
[6] Trevor J. Hastie, Robert John Tibshirani, and Jerome H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer series in statistics. Springer, New York, 2 editions, 2009.

**Abbreviations and notations**

CBG, census block group
GLM, generalized linear models
GBM, gradient boosting machine
ML, machine learning
SWOE, smooth weight of evidence

### Biography of the Author

**Jie DAI** is Associate actuary at Sentry Insurance Company in Middleton, WI. He is responsible for nonstandard auto modeling. He has a degree in aerodynamics from the Northwestern Polytechnic University in China. He is a Fellow of the CAS.