Why High Dimensional Modeling in Actuarial Science

Katrien Antonio and Simon CK Lee

Faculty of Economics and Business, KU Leuven, Belgium

July 10, 2014

KU LEUVEN

Presentation Outline

1. Introduction
2. Problem
3. Model Candidates
4. Data
5. Collision Frequency Models
6. Performance Assessment
7. Conclusion

Introduction

Predictive modeling has immense popularity in information analytics. For example, it can be used to

- derive cross-sales or up-sales opportunities
- vary prices by the time booked or by classes of the product for hotel or flight bookings
- analyse time spent on a website in order to better match what users need and what search engines output
- predict clinical results

In short, the ability of information analytics and predictive modeling to make accurate predictions has transformed society.

Predictive Modeling Techniques

Many predictive modeling techniques have been developed to fit the great variety of applications that have arisen.

Popular modeling choices include

- Support Vector Machines
- Boosting
- Random Forests
- Artificial Neural Networks
- Classification and Regression Trees

Each option differs in how the modeling problem is framed and how the prediction is derived.

But all these models aim to extract patterns between the explanatory variables and the response.

The real pattern is generally high dimensional.

Current Industry Trend for Information Analytics

According to the 2013 predictive modeling benchmarking suvery by Towers Watson

- 71 percent of North American personal insurers indicated that some form of predictive analytics are either in place or will be in place in the next year
- Only 67 percent indicated the same in 2012
- Number has been increasing over time
- Numbers are even higher in Europe due to more competitive operational environments

Current Industry Trend for Information Analytics

- While actuaries believe they are fully embracing the advanced modeling technologies, discussions in the insurance industry are still heavily biased towards the application of Generalized Linear Models (GLM).
- Publications on how to apply GLM for pricing, reserving, demand, economic capital models are numerous and techniques that help reduce extreme predictions from GLMs are popular topics at actuarial conferences.
- However, high dimensional techniques are not yet prevalent among actuaries.

Actuarial Bias

Several factors contribute to the actuarial bias towards GLM instead of higher dimensional techniques.

- Regulatory Resistance
- Varying levels of statistical training within stakeholder groups
- Operational limitation of the IT infrastructure

Data Mining

- There is tremendous value in exploring high dimensional modeling in the insurance industry
- Data mining techniques can significantly outperform GLM if there are strong interactions among variables
- Data mining can detect interactions, select variables and handle missing values simultaneously within the modeling process
- Leveraging features of data mining techniques will make the pricing process more efficient and effective

Presentation Outline

1 Introduction

2 Problem

3 Model Candidates

4 Data

5 Collision Frequency Models

6 Performance Assessment

7 Conclusion

Function Approximation

Actuarial ratemaking applications attempt to model key values such as

- Claim Frequency
- Claim Severity
- Conversion
- Retention

These models are classified as supervised learning, which means a *response* is to be predicted.

Data System

Mathematically, a system of data contains

- entries with response variables, y
- predictive co-variates, $\mathbf{x} = \{x_1, x_2, \ldots, x_k\}$.

The co-variates and response are assumed to be linked by an unobserved mapping F and a strictly monotonic *link function*. The goal is to find an estimate function F^* that minimizes a specified loss function $\Phi(y, F(\mathbf{x}))$, mathematically represented as

$$F^*(\mathbf{x}) = \underset{F(\mathbf{x})}{\operatorname{argmin}}\, E_{\mathbf{x}}[E_y(\Phi(y, F(\mathbf{x}))|\mathbf{x})] \tag{1}$$

Loss Function

Not every function is a loss function. A loss function should fulfill the following conditions.

Definition

A function, $\Phi(y, F(\mathbf{x}))$, is a loss function if it satisfies all the following conditions.

1. **Identifiable:** if $\Phi(y, F_1(\mathbf{x})) = \Phi(y, F_2(\mathbf{x}))\ \forall y$, $F_1(\mathbf{x}) = F_2(\mathbf{x})$.
2. **F-convex:** $\Phi(y, F(\mathbf{x}))$ is convex on $F(\mathbf{x})$ and is strictly convex at $F_{min}(\mathbf{x})$ where $F_{min}(\mathbf{x}) = \underset{F(\mathbf{x})}{\operatorname{argmin}} \Phi(y, F(\mathbf{x}))$.

   In the problem of function estimation, $F_{min}(\mathbf{x}) = g(y_i)$.
3. **Y-convex:** $\Phi(y, F(\mathbf{x}))$ is convex on y.
4. **Closed:** The set where $\Phi(y, \cdot)$ is defined is closed.

Loss Function

- The prediction $\hat{y}$ is equal to $g^{-1}(F^*(\mathbf{x}))$.
- Most commonly used loss function in actuarial predictive modeling is deviance
- D, is a negative linear transformation of loglikelihood, ll

$$D = -2ll + C \tag{2}$$

  where C is a constant
- Thus, minimizing the deviance is equivalent to maximizing the log-likelihood.

Presentation Outline

1 Introduction

2 Problem

3 Model Candidates

4 Data

5 Collision Frequency Models

6 Performance Assessment

7 Conclusion

Generalized Additive Models

- Covers many existing solutions to the function approximation problem
- Mathematically, the class consists of an algorithm represented in the form:

$$F^*(\mathbf{x}) = g(\sum_{j=1}^{J} f_j(\mathbf{x})) \qquad (3)$$

- Each family member has its unique way to specify the basis function and assign parameters
- Basis function and parameters are combined to form $f_j(\mathbf{x})$

Members of the Generalized Additive Model Family

- Generalized Linear Models
- Generalized Additive Models
- Artificial Neural Network
- Gradient Boosting Machine
- Delta Boosting Machine
- Classification and Regression Tree

Gradient Boosting Machine

- Gradient descent approach with numerical optimization and statistical estimation
- Iteratively minimize the loss function between the actual observation and the corresponding prediction
- Top-tier predictive model among data mining techniques
- Simplicity of the algorithm is inspiring more research that could lead to even more powerful extensions

$$F^*(\mathbf{x}) = \sum_{t=1}^{T} f_t^*(\mathbf{x}) = \sum_{t=1}^{T} \beta_{\mathbf{t}} h(\mathbf{x}; \mathbf{a}_t) \qquad (4)$$

Algorithm

Algorithm 1 Forward Stagewise Additive Modeling

```
Initialize $F_0(\mathbf{x})$
for $t = 1$ to $T$ do
    Estimate $\beta_{\mathbf{t}}$ and $\mathbf{a_t}$ by minimizing $\sum_{i=1}^{N} \Phi(y_i, F_{t-1}(\mathbf{x}_i) + \beta_{\mathbf{t}} h(\mathbf{x}_i; \mathbf{a}_t))$
    Update $F_t(\mathbf{x}) = F_{t-1}(\mathbf{x}) + \beta_{\mathbf{t}} h(\mathbf{x}; \mathbf{a}_t)$
end for
Output $hatF(\mathbf{x}) = F_T(\mathbf{x})$
```

Delta Boosting Machine

- Modification to GBM that better utilizes the base learner
- Primary difference between DBM and GBM is their sorting rules
- In GBM, the gradient of each observation is used as the sorting element:

$$r_i = -\left[\frac{\partial \Phi(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F_{t-1}(\mathbf{x})}, \ i = \{1, \dots, M\}$$

- DBM attempts to reduce the deviance to the maximum extent at each iteration and thus the minimizer also called delta is used in sorting:

$$\delta_i = \underset{s}{\operatorname{argmin}}\, \Phi(y, F_{t-1}(\mathbf{x}_i) + s), \ i = \{1, \dots, M\}$$

Presentation Outline

1 Introduction

2 Problem

3 Model Candidates

4 Data

5 Collision Frequency Models

6 Performance Assessment

7 Conclusion

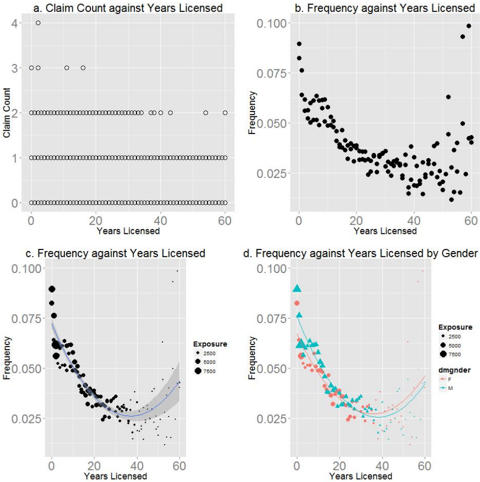
Function Approximation

- Real-life data from a Canadian insurer
- Consists of policy and claim information at the vehicle level for Collision coverage in a particular province
- Includes the experience for calendar/accident years 2001 to 2005
- Response to be predicted is the claim frequency
- 290,147 earned exposures and an overall claim frequency of 4.414%
- Imbalanced or skewed class distribution for the target variable
- Rigorous exploratory data analysis and variable selection process are more influential to the outcome of the modeling

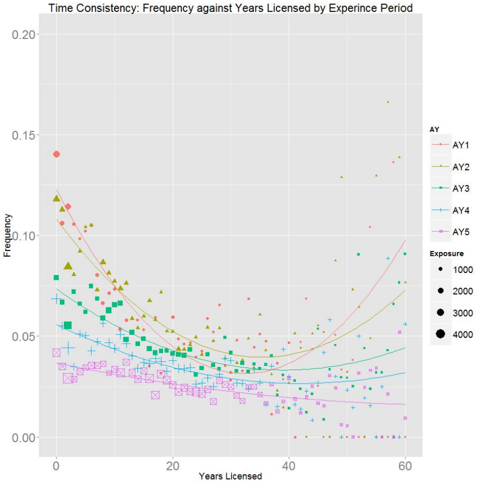
Exploratory Data Analysis

- Willingness to look for what can be seen
- No set rules on how to formalize the process
- Encouraged to consult experts from underwriting, sales, IT and claims
- Visual tools required are dependent on the problems and style of actuaries
- Following tools are handy for typical personal auto pricing purposes

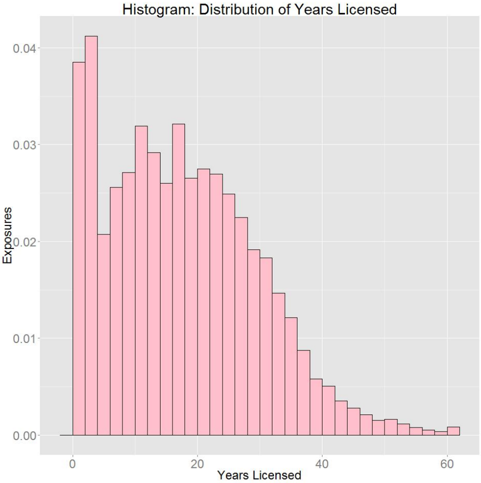
Scatter Plots



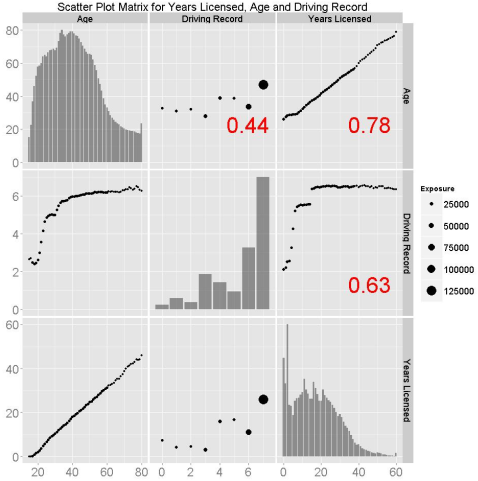
Time Consistency Plot



Histograms


Histogram: Distribution of Years Licensed
Exposures
0.04
0.03
0.02
0.01
0.00
0
20
40
60
Years Licensed

Correlation Plots



Principal Components

- Many explanatory variables are correlated in auto pricing
- Principal components are created to transform elements into linearly uncorrelated variables
- Done through iterative eigen decomposition
- Compress the list of variables by dropping variables that do not significantly explain the variability of data

	PC 1	PC 2	PC 3
Years Licensed	-0.67	0.73	0.12
Driving Record	-0.07	0.11	-0.99
Age	-0.74	-0.67	-0.02
Variability Captured	0.88	0.11	0.01

Preliminary Variable Selection

- Good EDA gives actuaries a clear picture about the data and which variables should be selected for modeling
- A predictive variable usually exhibits a clear pattern against the response
- Pattern need not to be straight line or monotonic
- Robust selection should leave actuaries with only a handful of variable combinations for modeling
- Leaving too many options to modeling stage will significantly lengthen the modeling process and likely result in overfitting

Data with Unusual Values

- Real-life data is seldom perfectly clean
  - inaccurate information from the insured
  - omission from the agents
  - system errors
  - mismatch from external source
- Need a robust process to validate the quality of data
- Many mistakes can be easily spotted
  - negative years licensed
- Other checks include, but are not limited to
  - the number of vehicles insured to confirm the discount on insuring multiple vehicles
  - the age of vehicle when it was purchased to confirm the model year is accurate
  - the number of years insured to confirm the years licensed
  - the address to confirm the territory of the insured

Missing Data

- Some predictive modeling tools require all fields to be complete
- Common rules of thumb that attempt to solve the problem:
  - Replacing the missing values with the field average
  - creating a new variable that indicates the value for the observation is missing
  - finding a proxy to approximate the value from other variables
  - deleting the observations
- Around 1200 observations (less than 0.15%) have missing fields in our data
- Observations are deemed insignificant and deleted

Presentation Outline

1 Introduction

2 Problem

3 Model Candidates

4 Data

5 Collision Frequency Models

6 Performance Assessment

7 Conclusion

Frequency Models

- Use GLM, GAM, ANN, GBM, DBM and CART as the candidates for the frequency model
- Partitioned the data into train (80%) and test (20%) data sets through random sampling
  - Train data is used for modeling
  - Test data is used as an independent source to verify the performance
- For models that requires intermediate validation data, the train data is further split into pure train (80%) and validation (20%) of the train data
- Ex-ante belief that claim count follow the Poisson distribution
- Poisson deviance is used as the basis of performance
- Best model should have lowest deviance

Initial Run

Model	Train Deviance
GLM	0.00
GAM	-359.78
GBM	-882.77
DBM	-897.27
ANN	103.79
CART	2095.85

- For the difference between GLM and DBM to be considered to be statistically immaterial at 5% significance, DBM has to have at least 850 more parameters than GLM.
- For the case between GLM and GAM, GAM has to have at least 318 more parameters.

Presentation Outline

1 Introduction

2 Problem

3 Model Candidates

4 Data

5 Collision Frequency Models

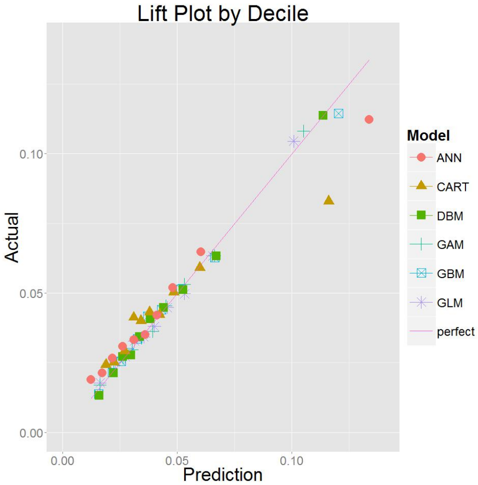
6 Performance Assessment

7 Conclusion

Performance Assessment

- Assess the performance of the candidates by utilizing the lift plot
- Sort the prediction and group the observations into 10 deciles
- Lift is defined to be the ratio of the mean of the response in the top decile and the mean of the bottom decile
- A high lift implies the model's ability to differentiate observations
- If points are aligned with the line $y = x$, the model has a high predictive performance

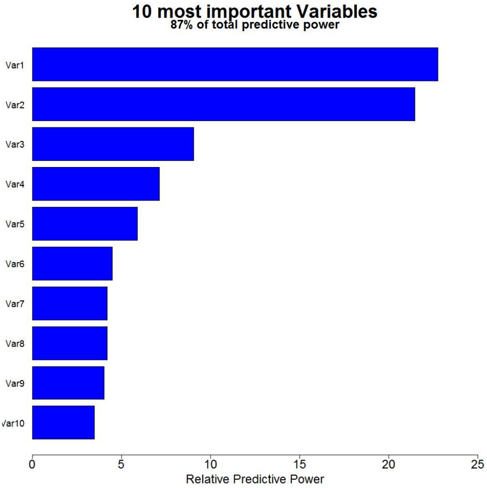
Lift Plot


Lift Plot by Decile
Actual
Prediction
0.00
0.05
0.10
0.00
0.05
0.10
Model
ANN
CART
DBM
GAM
GBM
GLM
perfect

Variable Importance

- Focus on which variables exert more influence on the model performance
- Likelihood improvement at each iteration is assigned to the each variable and improvements of all iterations are then aggregated
- Importance is normalized such that the sum equals 100 for easier comparison

Variable Importance

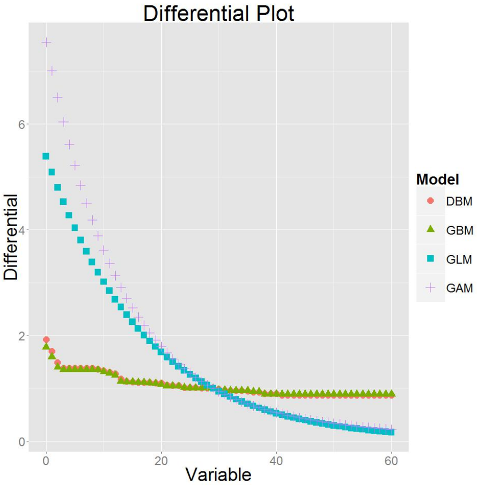


Comparison of Variable Importance between Models

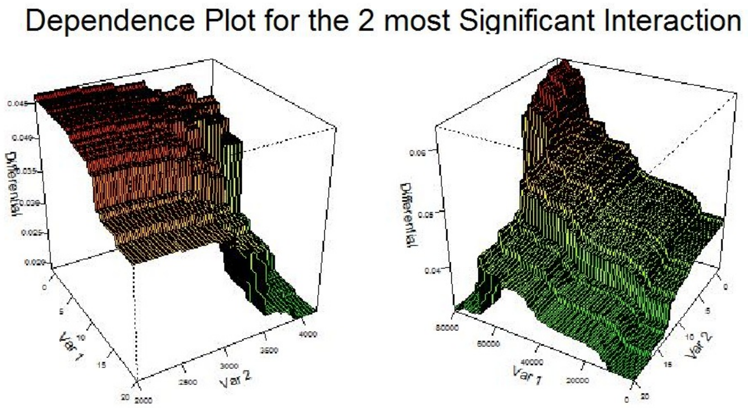
Variable	Importance	GLM Rank
Var1	22.80	7
Var2	21.50	1
Var3	9.10	20
Var4	7.10	3
Var5	5.90	2
Var6	4.50	15
Var7	4.20	11
Var8	4.00	6
Var9	3.50	13
Var10	3.10	9

- Great opportunity to improve the GLM models by comparing the differential plots of those variables that show vastly different rankings between models

Marginal Differential Plot



Joint Differential Plot



Dependence Plot for the 2 most Significant Interaction

Final Run

- Magnitude of improvement decreases as we drill down to the less significant variables or interactions
- Over-fitting may also result when one attempts to temper the model too much
- Consider reviewing 5 to 10 most significant variables and including 2 to 4 interactions
- Several iterations of revision may be necessary
- Final model have considerably fewer parameters than the initial one due to removal of many highly correlated variables to reduce the co-linearity effect

Results on Holdout Data

- Test data is the one and only independent benchmark in comparing all candidates
- Should not be accessed until results of all the model candidates are final

Model	Train Deviance	Test Deviance
GLM	0.00	0.00
GAM	-359.78	-70.92
GBM	-882.77	-147.46
DBM	-897.27	-165.73
ANN	103.79	9.87
CART	2095.85	514.28
GLM Final	-82.93	-30.00

Results on Holdout Data

- For the improvment to be deemed as statistically insignificant at 5%, the finalized version has to have 70 more parameters than the original one
- Confirms the superiority of the finalized model
- Conclude the modeling with one last re-run using the full set of data to maximize the utility of the data

Presentation Outline

1 Introduction

2 Problem

3 Model Candidates

4 Data

5 Collision Frequency Models

6 Performance Assessment

7 Conclusion

Conclusion

- Provide a simplified framework for actuarial pricing
  - data cleaning
  - exploring
  - modeling
- Data quality is the key to the success of actuarial pricing
- Requires actuaries to have assess to various functional experts within the their organizations
- Many tools are available to actuaries to visualize the interdependence among data at no cost
- Inference from other models can help actuaries to tailor the pricing algorithm to best describe the data behavior

Conclusion

- Modeling procedure is usually recursive
- Once comfortable with the finalized model, actuaries should verify the results with an independent holdout data
- Modifications may be necessary to reflect the nature of different problems
  - Contour plot for analysis of rating territories
  - Compound modeling where the geo-spatial residual is used for spatial smoothing
  - Exposure rating or increase limit factor analysis when new deductibles or limits are introduced