Regression Project

Fall 2014

Tony, Liu

Predict expected fire losses for insurance policies

A Fortune 100 company, Liberty Mutual Insurance has provided a wide range of insurance products and services designed to meet our customers' ever-changing needs for over 100 years.

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.

Because we seek to tap innovation both inside and outside the company, certain eligible Liberty Mutual employees are encouraged to participate in this challenge for development purposes. Refer to the competition rules for the full details.

Started: **12:00 pm, Tuesday 8 July 2014 UTC**
Ends: **11:59 pm, Tuesday 2 September 2014 UTC (56 total days)**

Data Files

File Name	Available Formats
sampleSubmission.csv	.zip (687.60 kb)
test.csv	.zip (553.31 mb)
train.csv	.zip (554.04 mb)

This data represents almost a million insurance records and the task is to predict a transformed ratio of loss to total insured value (called "target" within the data set). The provided features contain policy characteristics, information on crime rate, geodemographics, and weather.

The train and test sets are split randomly. For each id in the test set, you must predict the target using the provided features.

Data Fields

id : A unique identifier of the data set

target : The transformed ratio of loss to total insured value

dummy : Nuisance variable used to control the model, but not working as a predictor

var1 – var17 : A set of normalized variables representing policy characteristics (**note: var11 is the weight used in the weighted gini score calculation**)

crimeVar1 – crimeVar9: A set of normalized Crime Rate variables

geodemVar1 – geodemVar37 : A set of normalized geodemographic variables

weatherVar1 – weatherVar236 : A set of normalized weather station variables

Data Type

Categorical Variable Name	Variable Type	Possible Values

var1	Ordinal	1, 2, 3, 4, 5, Z*
var2	Nominal	A, B, C, Z*
var3	Ordinal	1, 2, 3, 4, 5, 6, Z*
var4+	Nominal	A1, B1, C1, D1, D2, D3, D4, E1, E2, E3, E4, E5, E6, F1, G1, G2, H1, H2, H3, I1, J1, J2, J3, J4, J5, J6, K1, L1, M1, N1, O1, O2, P1, R1, R2, R3, R4, R5, R6, R7, R8, S1, Z*
var5	Nominal	A, B, C, D, E, F, Z*
var6	Nominal	A, B, C, Z*
var7	Ordinal	1, 2, 3, 4, 5, 6, 7, 8, Z*
var8	Ordinal	1, 2, 3, 4, 5, 6, Z*
var9	Nominal	A, B, Z*
dummy	Nominal	A, B

* : Level "Z" in these variable represents a missing value. Missing values elsewhere in the data are denoted with NA

+: Levels for var4 are in a hierarchical structure. The letter represents higher level and the number following the letter represents lower level nested within the higher level.

Numeric Variable Name	**Variable Type**
target	Continuous
id	Discrete
var10	Continuous

var11	Continuous
var12	Continuous
var13	Continuous
var14	Continuous
var15	Continuous
var16	Continuous
var17	Continuous
crimeVar1 – crimeVar9	Continuous
geoDemVar1 – geoDemVar37	Continuous
weatherVar1 – weatherVar236	Continuous

SOLUTIONS

1 decrease the numbers of variables

The CrimeVars include 9 variables, the GeodemVars include 37 variables, and the weatherVars include 236 variables. They are too many to fit the model. As they are all quantitive variables, I use the *Principle Components Analysis Method* to decrease the numbers of them using R. As a result, I transform the 9 CrimeVars into 2 new variables, 37 GeodemVars into 3 new variables, and 236 WeatherVars into 8 new variables. They are

Variable Name	Variable Type
CrimeNewVar1	Continuous
CrimeNewVar2	Continuous
GeodemNewVar1	Continuous
GeodemNewVar2	Continuous
GeodemNewVar3	Continuous
WeatherNewVar1	Continuous
WeatherNewVar2	Continuous
WeatherNewVar3	Continuous
WeatherNewVar4	Continuous
WeatherNewVar5	Continuous
WeatherNewVar6	Continuous
WeatherNewVar7	Continuous
WeatherNewVar8	Continuous
WeatherNewVar9	Continuous
WeatherNewVar10	Continuous

2 fit the severity model

There are 1188 samples whose response variables — Target — are not zero. These samples construct severity model's data. Figure 1 shows the histogram of the distribution of Target. It is right-skewed seriously. Figure 2 shows the scatterplot of Target and Var11. It seems that Var11 is also right-skewed and the relationship with Target and Var11 is negative correlate.

I use GLM with Gamma distribution and log link function. It seems that Var11 is the denominator of Target, so I transform Var11 into log form. Figure 3 shows the scatterplot of Target and log Var11.

Figure 1

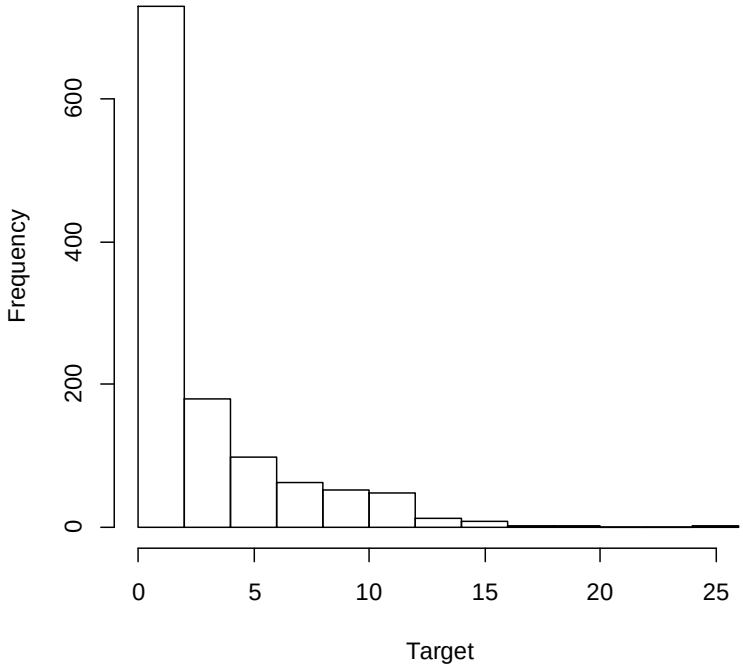


Figure 2

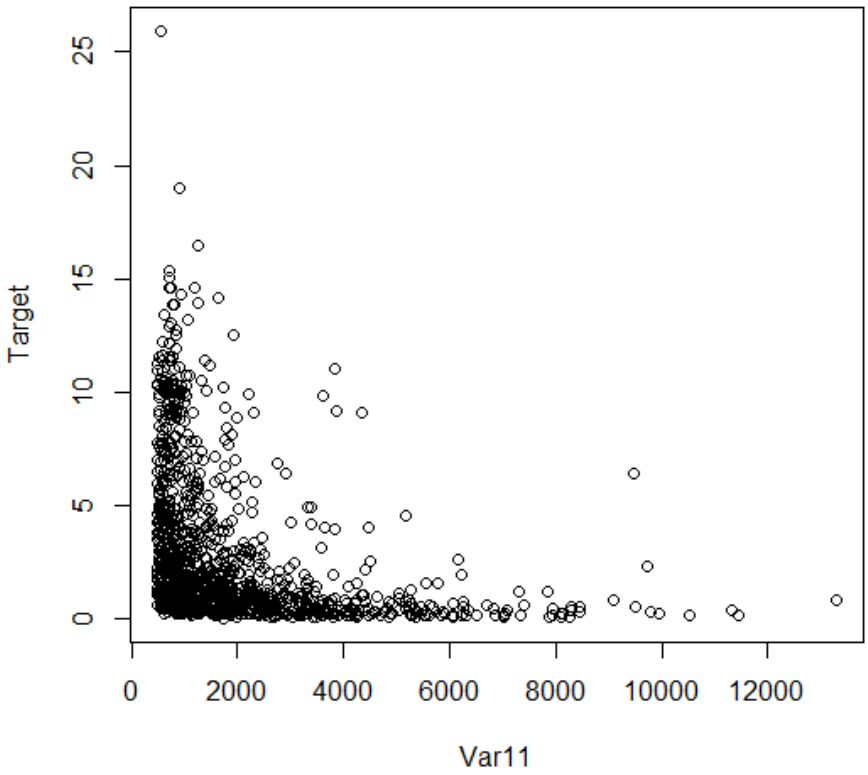


Figure 3

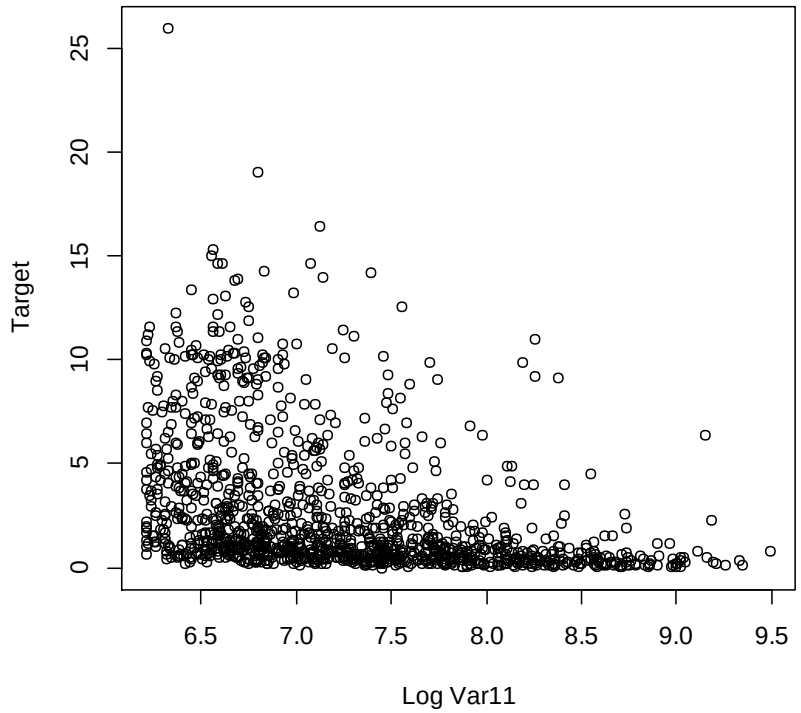
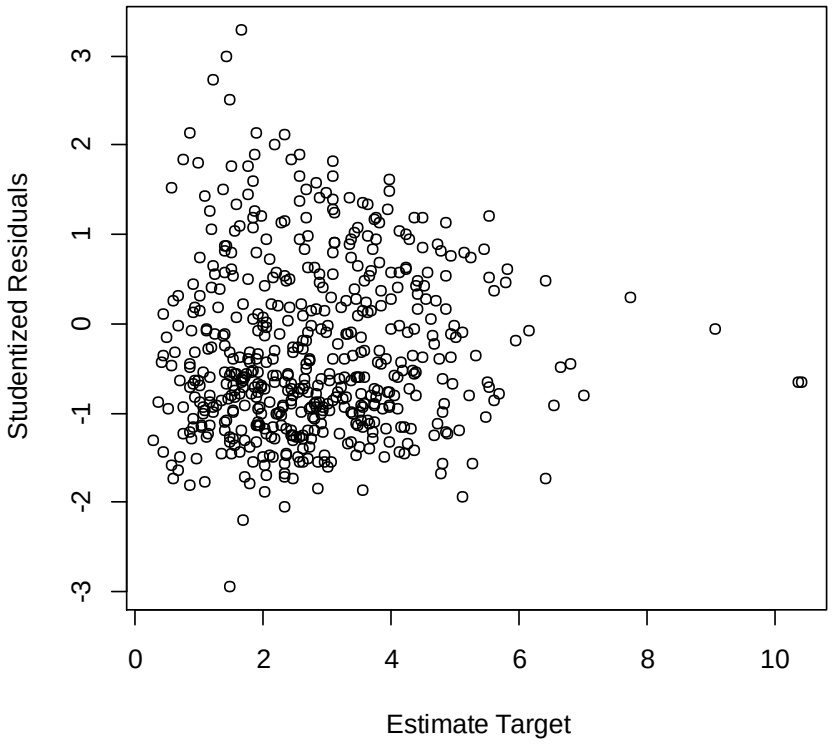


Figure 4



Stepwise the regression using R, I get the optimal severity model. Table 1 shows the estimated parameters. Figure 4 shows the scatterplot of estimated Target and studentized residuales.

Table 1: Severity Model

	Estimate	Std Error	T Value	Pr(>|t|)
(Intercept)	1.5904	1.5034	1.058	0.2906
LogVar11	-0.8772	0.0933	-9.399	< 2e-16
VAR7	-0.0562	0.0338	-1.662	0.0971
VAR9B	-0.2537	0.1376	-1.844	0.0657
VAR10	1.3879	0.4058	3.42	0.0007
VAR13	-0.1372	0.0788	-1.741	0.0822
VAR15	0.0045	0.0019	2.347	0.0193
WeatherNewVar10	0.0365	0.0212	1.716	0.0867
CrimeNewVar1	-0.0966	0.0597	-1.619	0.1061

3 fit the frequency model

There are more than 600,000 samples, R is no longer satisfied, I use *Emblem* to fit the frequency model. Because the explanatory variables inputted in Emblem must be factors, I have to group the quantitive variables firstly.

Most of the sample's Target are zero. No zero Target means there was a claim, while zero means there wasn't. So I use Logit Model.

Figure 5 shows the relationship between grouped Var11 and its estimated parameters. It looks like a concave curve. So I transform grouped Var11 into an orthogonal curve with 2 powers.

Stepwise the regression, I get the optimal frequency model. Table 2 shows the estimated parameters. Figure 6 shows the gain curve of the model, its Gini coefficient is 0.4438.

Figure 5

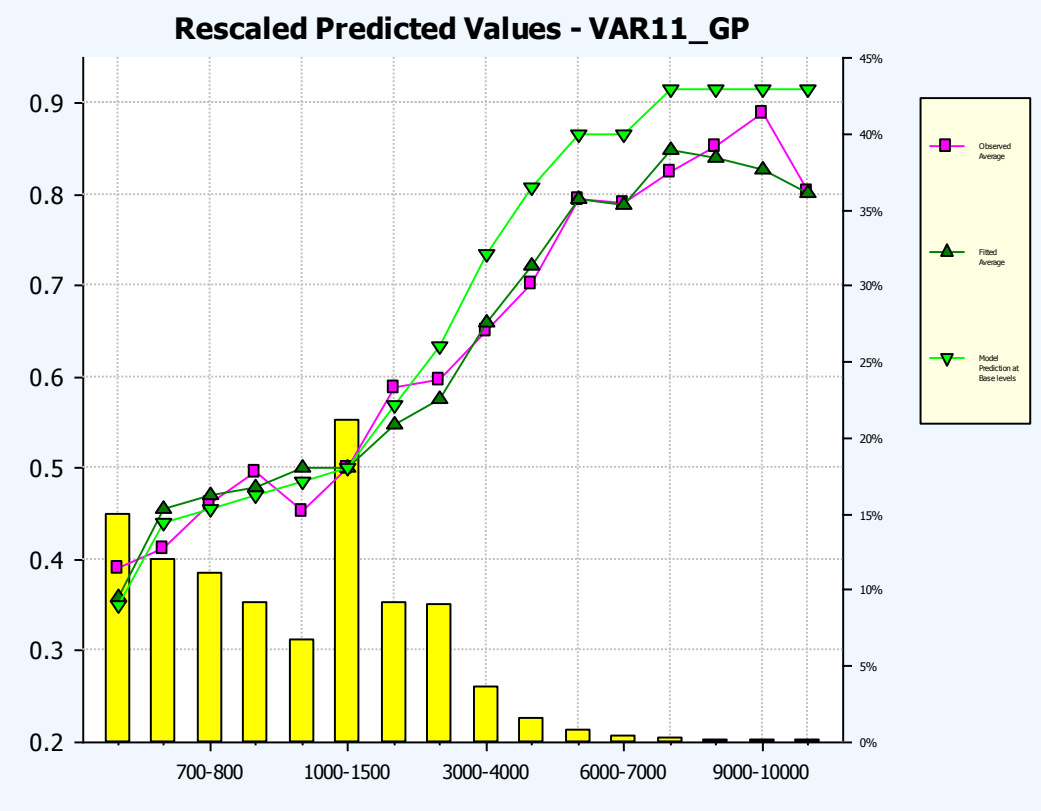


Figure 6

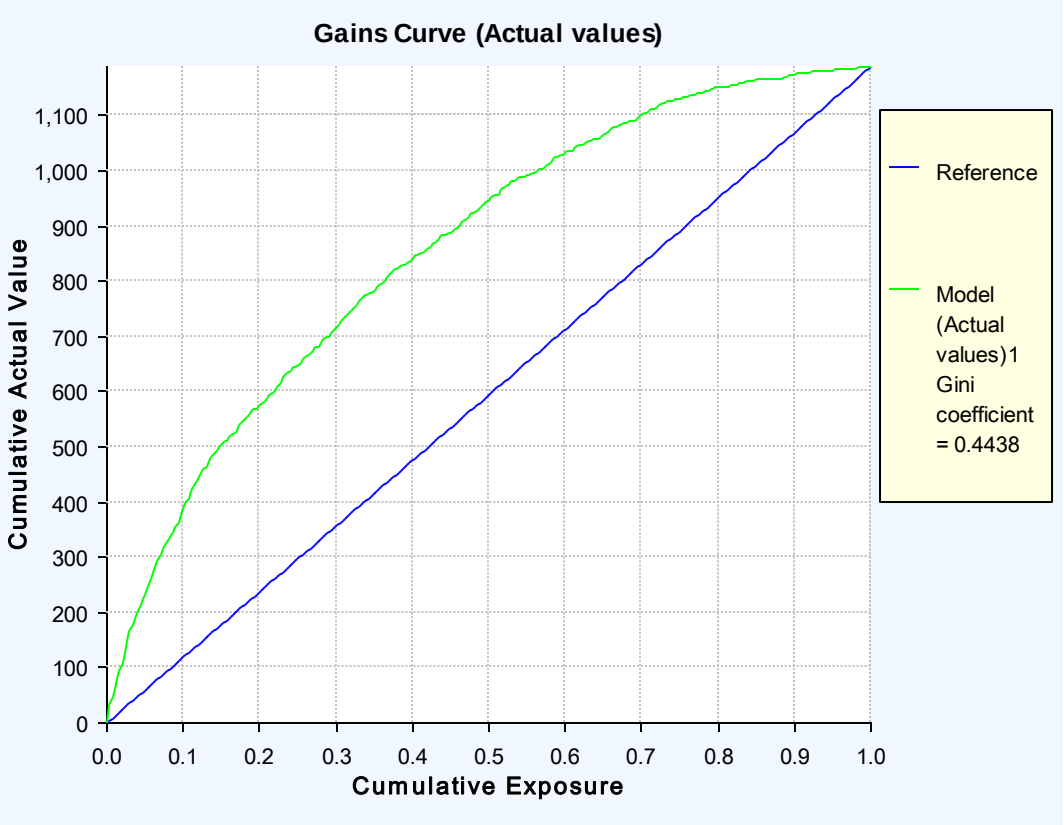


Table 2

		Value	Std Error	Chi-Squared Test
(Intercept)		-6.1921	0.1452	0.00%
DUMMY	(A)			2.50%
	(B)	0.0616	0.12834	
Var8_GP	(1)	0.2572	0.07912	0.00%
	(2)			
	(3,4)	-0.1609	0.07751	
	(5,6)	-0.3528	0.1497	
Var10_GP	(<4)			0.00%
	[4,4.2)	-0.2683	0.07756	
	[4.2-4.4)	-0.3417	0.1363	
	[4.4-4.6)	-0.9386	0.18942	
	[4.6,+)	-1.5539	0.22582	
Var13_GP	(0)	0.411	0.0915	0.00%
	(0,1.5)	0.1981	0.08262	
	[1.5,2)			
	[2,+)	-0.3054	0.12625	
Var15_GP	(0,10)	-0.1283	0.10024	0.06%
	[10,30)			
	[30,60)	0.1904	0.07353	
	[60,90)	0.3	0.11088	
	[90,150)	-0.1995	0.19411	
	[150,+)	-0.521	0.57462	
GeodemNewVar2_GP	(<-10)	0.0718	0.12271	0.00%
	[-10,-1.5)	-0.2948	0.0679	
	[-1.5,+)			
WeatherNewVar2	(<2)			0.00%
	[2,5)	0.1031	0.10415	
	[5,10)	0.2095	0.10724	
	[10,+)	-0.037	0.1255	
WeatherNewVar3	(<-3)	0.0006	0.10457	0.42%
	[-3,-1)			
	[-1,0)	0.0683	0.10666	
	[0,+)	-0.2608	0.11907	
Var4_GP	(A-D,E5-H,N1,O)			0.00%
	(E1-4,P+)	0.3959	0.08412	
	(I-M)	0.8515	0.07094	
Var11_GP	OPoly(1)	0.9534	0.05388	0.00%
	OPoly(2)	-0.141	0.03142	