Report of Logistic Regression with $l_1$ Penalty Method on Insurance Claim Project

Meng Wang
March 11, 2013

1 Goal of the project

We would like to build a model to predict whether a claim belongs to a fast track defined as the total charge will be less than 5000 dollars based on the information collected from the first three days after the accident occurs. We expect our model has relatively high capability of prediction and interoperation. From the benefit of the insurance company, we will select first 20 significant predictors to make a questionnaire as the basic information to collect just after the accident.

The following is from Iris.

Today most insurance claims are triaged and routed in the same manner regardless of the characteristics of the claims. The objective of this project is to create a model that can be used to predict the likelihood of an auto casualty claim falling into the low severity group. If a claim is being identified with high probability of low severity, it can be routed to cost effective channel, i.e. Fast Track, for processing. If the probability of low severity is low, that means more caution is required in processing the claim. The severity is referred to the total medical treatment charges. The low severity group is being defined as the total medical treatment charges less than $5,000. The high severity group is being defined as the total medical treatment charges greater than $5000. The threshold of $5,000 is set for research purpose. The model can be adjusted with different value based on real world scenario. By applying this model to claim triage workflow, it will result in cost saving, adjuster productivity increase, and workflow efficiency for insurance carriers.

One big challenge is to find the balance between accuracy and number of parameters in the final model. Since the parameters of the model will be collected in the form of asking a questionnaire through phone call to the claimant or medical providers, it is important to control the numbers of questions that to be asked. 20 questions would be the maximum number questions. Secondly, the data has high dimensions, making the data very sparse. Therefore, it is more difficult to find signals out of noise. Last but not the least, since the $5,000 cut is a business decision. The cases with total medical charged amount slightly lower or higher than the cut $5,000 are not very informative.

2 Data Description

There are about 130,000 claims (observations), 1300 features (predictors) and the sparsity of the data is an issue. From Iris’ description of data, the features include

- Characteristics of vehicle, such as vehicle make, model, year, etc.
- Characteristics of car accidents, such as airbag deployment status, area of vehicle damages, point of impact, cost to repair vehicle, etc.
• Diagnosis and medical treatment, such as type of treatment, cost, duration, and sequence
and treatment, etc.

This data has gone through data preparation, cleansing, and been transformed into meaningful
indicators and measurements. These cases were derived from the matched claims between 2.5M
eligible auto casualty claims and 11M auto collision estimates from 2009 to 2011. Each auto
casualty claim was comprised of 60 treatment details on average.

And there are a lot of binary variables such as medical treatments and non 0-1 variables are:
ClaimantID, body.part.count, DMGCount, AgeAtLossAdj, AgeAtLoss, VehicleAge, ALL-FRAME-
LABOR-HRS, ALL-FRAME-LABOR-AMT.

3 Prepare data for the model

Since we would like to build the model based on first three days, we focus on the variables collected
only from first three days. For the goal of interpretation, we consider the correlation between
predictors. We use the function corr() in R to get pairwise correlation between variables. (We try
to use variance inflation factor to inference correlated variables, but the function vif() doesn’t work
probably because our data includes both numerical and categorical variables.) After combining
correlated variables, we finally get 664 predictors. And we normalize the numerical variables such
that they have zero mean and 1 standard deviation. Later we will fit logistic regression and penalize
\( l_1 \) norm of the coefficients so that it is important to normalize the predictor variables so they are
unit-less.

We set response as 0-1 binary categorical variable; if the total charge is greater than \$5000, the
response is 1; otherwise is 0. And therefore this is a classification task.

4 Fit logistic regression with \( l_1 \) penalty model

As Iris said, they fit logistic regression before but they didn’t satisfy the prediction ability of this
model. As a review, here is the logistic model.

Assume the response \( y \) is binary, e.g., \( y \in \{0, 1\} \). We model \( \mu(X) = P(y = 1|X) = E(y|X) \) with
\[
\text{logit}(X) := \log \left( \frac{\mu(X)}{1 - \mu(X)} \right) = \beta_0 + \beta_1^T X,
\]
and fit the model by maximum likelihood, i.e., choosing \( \beta = (\beta_0, \beta_1) \) such that maximizing the likelihood
\[
\sum_{i=1}^{n} y_i \log(\mu_i) + (1 - y_i) \log(1 - \mu_i) = \sum_{i=1}^{n} y_i \log(g^{-1}(\beta_0 + \beta^T x_i)) + (1 - y_i) \log(1 - g^{-1}(\beta_0 + \beta^T x_i)),
\]
where \( g = \text{logit} \).

To focus on the signals—the subset of data relatively strongly representing the class labels, we
will use the data with total charge \( > 6000 \) and \( < 4000 \) to build the model.