Will your TSA claim be approved? - A predictive modeling approach

Tingting Sun, Jianguo Jiang, Jiamin Zhao, Supervisor: Dr. Lu Xiong

Actuarial Sciences, Department of Mathematical Sciences. Email:ts7f@mtmail.mtsu.edu

Abstract

The air passengers claim for the baggage damaged, lost, and delayed in transit is a significant problem for the airport. This article presents a model with a high AUC (Area Under ROC) that can be used to predict the rate of bagage claim denied. The data set of claims against the Transportation Security Administration (TSA) from 2008 to 2010 includes claim date, type, site, claim amount, and disposition as well as airport code and airline name. We used R to process data and applied the clustering method, LASSO regression, and cross-validation method as well as regularization to predict the model. The AUC of the GLM model that was used is promising. The variable importance plot is provided below.

Decision Tree

There are 5 levels of the variable Claim Type. Figure 3 shows the data set of claims against the Transportation Security Administration (TSA) from 2008 to 2010 includes claim date, type, site, claim amount, and disposition as well as airport code and airline name. We used R to process data and applied the clustering method, LASSO regression, and cross-validation method as well as regularization to predict the model. The AUC of the GLM model that was used is promising. The variable importance plot is provided below.

Generalized Linear Model

Then we considered the binomial Generalized Linear Model (GLM) with logit link function using regularization. The three types of regulation methods tested are LASSO, Ridge regression, Elastic Net. The three regulations are not much different in AUC (0.86–0.87), but LASSO has the potential advantage of removing more feature to build a simple model. Hence, we choose LASSO.

The cutoff value is the probability threshold above which the prediction is classified as Approve, below which it is classified as Deny. Since the estimation of cutoff values can change the classification results, we ran the GLM at different cutoff to obtain the best cutoff that achieves the maximum profit using the assumption provided by the market department about the cost and profit. When cutoff=0.23, we can maximize the expected profit, which is 32195.

A Model Demo for Marketing

The following table contains 1 base case and 7 variation cases from the base case. To do a prediction, we need 7 pieces of information in the input, including Report Lag, AirPort Code, Claim Amount etc. We use them to predict weather the TSA will deny or approve this case. The model used is the GLM selected in Task 8 with the cutoff=0.23 to maximize the total profit. The algorithm returns the probability of approve, and the probability above 0.23 is classified as Approve, otherwise classified as Deny. The probability and prediction results are listed in the last two column of the table. The probability is calculated using the coefficients table in Task 8 multiply the 7 variables in the input (after necessary preprocessing including binarization). This can be implemented in an Excel spreadsheet that is easy for the marketing department to use.