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

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

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

Abstract

The airline 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 baggage 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. We provided the comparison among the classification tree, random forest(RF), and generated linear model(GLM) in different aspects corresponding to the goodness of prediction. We chose GLM as the final model because of its great interpretability and high accuracy. The GLM can be implemented in an Excel spreadsheet that is easy to use, while it is difficult to do so for the other models.

Problem Statement

TSA is the acronym for the Transportation Security Administration, an American governmental agency that is responsible for travel safety, especially air travel. "You may file a claim if you are injured or your property is lost or damaged during the screening process."-- TSA.

If you are injured or your property get lost or damaged, you can file a claim with the TSA for reimbursement. Then after investigating your situation, which may spend up to 6 months, the TSA will either approve your claim and reimburse you the full amount, or deny your claim altogether. Historically, up to 2 in 3 claims got denied and do not get reimbursed. So you may ask -- Will your TSA claim be approved?

Data Preprocess

The raw data set has 7 columns of predictors and one column of the target variable, which is Deny or Approve. The predictor columns include the Report.Lag, Airport.Code, Airline.Name, Cliam.Amount etc. There are total 24628 cases in which 7270 are Approve cases and 17368 Deny cases. We replaced the Approve/Deny with a 0/1 target variable called value_flag.

We examined each predictor variable on its own and with respect to value_flag. The following figure is the distribution of Report.Lag, which is right skewed.

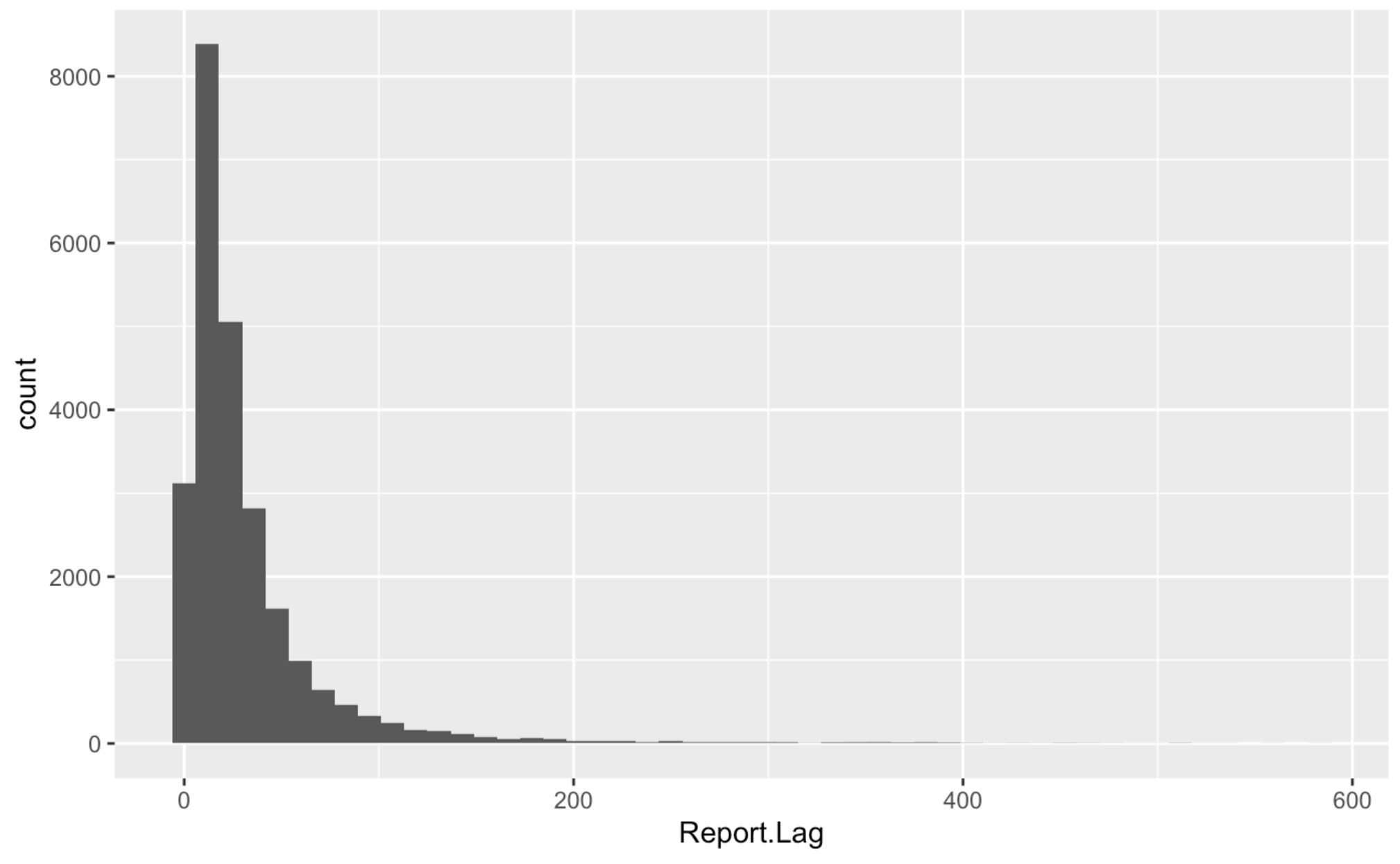

Figure 1: The histogram of Report Lag. It follows a right skewed distribution.

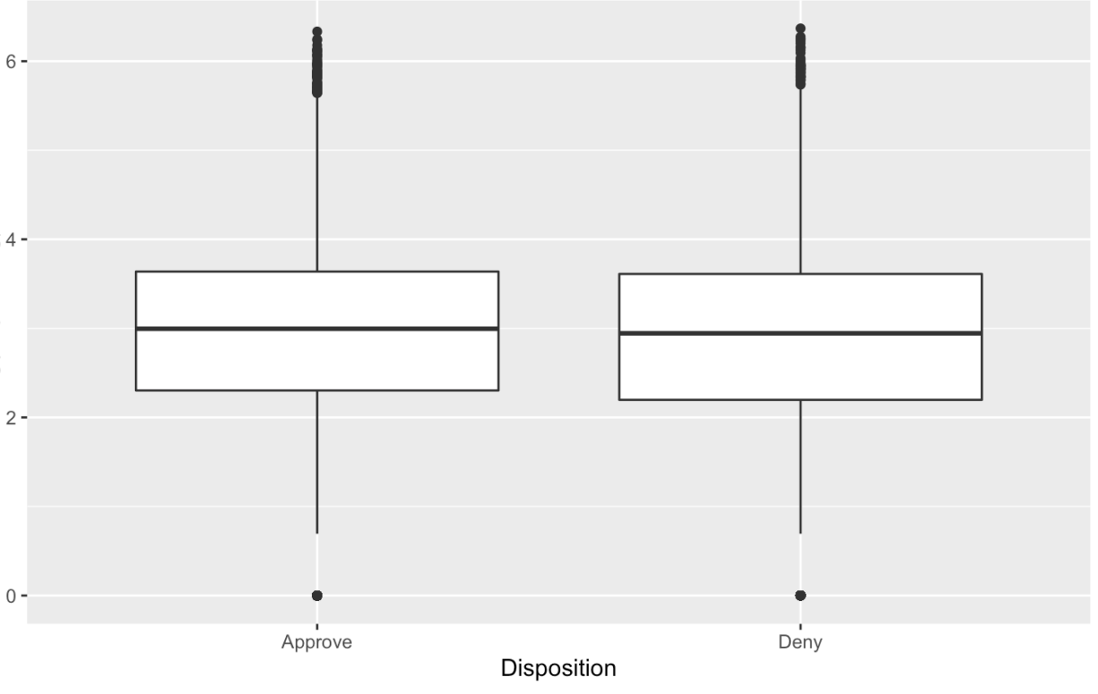

Figure 2: The boxplot by denied claims group and approved claims group.

We suspect the longer Report Lag could lead to higher deny rate. The boxplot below turns out there is no significant difference for the value of Report Lag days among denied claims and approved claims.

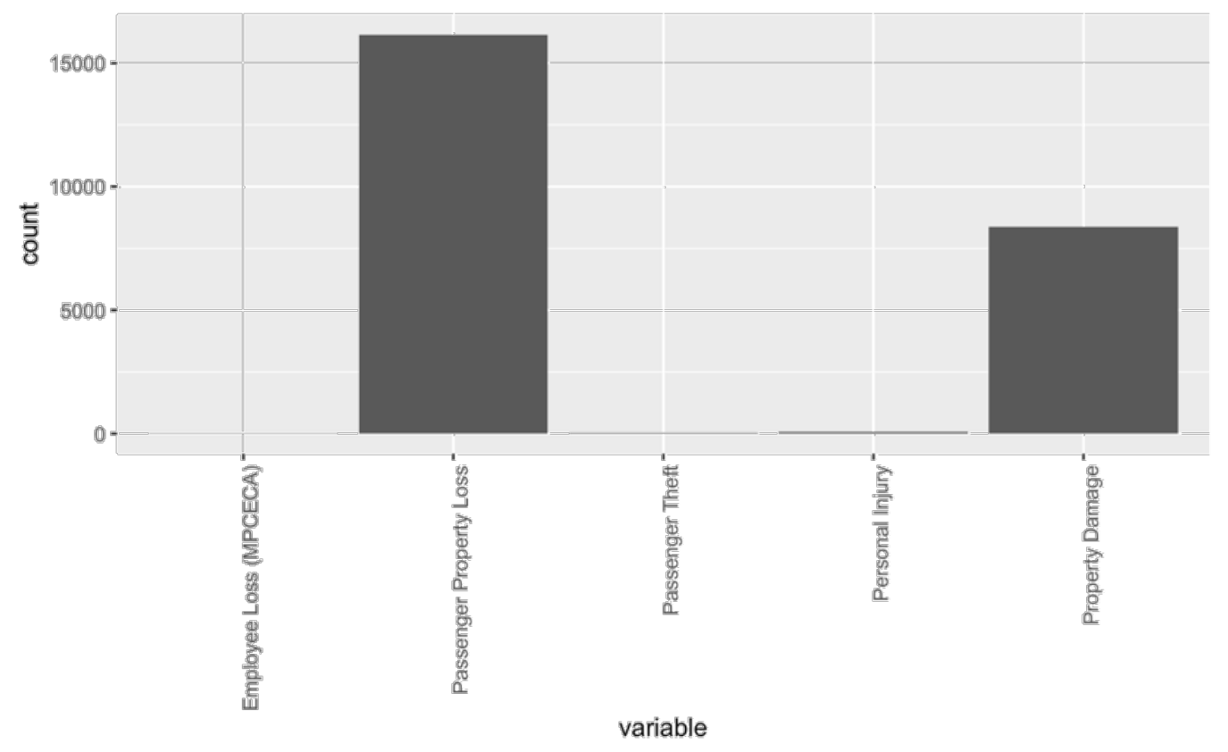

Figure 3: The boxplot of Claim.Type.

There are 5 levels of the variable Claim.Type. Figure 3 shows majority of the of claims are in two categories: Passenger Property Loss, Property Damage.

There are a lot of levels in each of the other three factor variables: Airport.Code, Airline.Name, Item. To improve the predictive power of these variables, we need reduce their levels. To do so, we use the k-means clustering (k=5). The levels with similar approve ratio are clustered together. In this way, we have reduced the levels to 5 for each of these variables.

Decision Tree

In cross validation (CV) method, a grid search was applied on a list of parameters. The data is split into k folds, with k - 1 folds is the training data, the other 1 fold is left as the validation data. The algorithm will try each value of parameter in the Grid list to do a k fold CV. The method 2 produced the same tree as method 1. The AUC of this tree is 0.78.

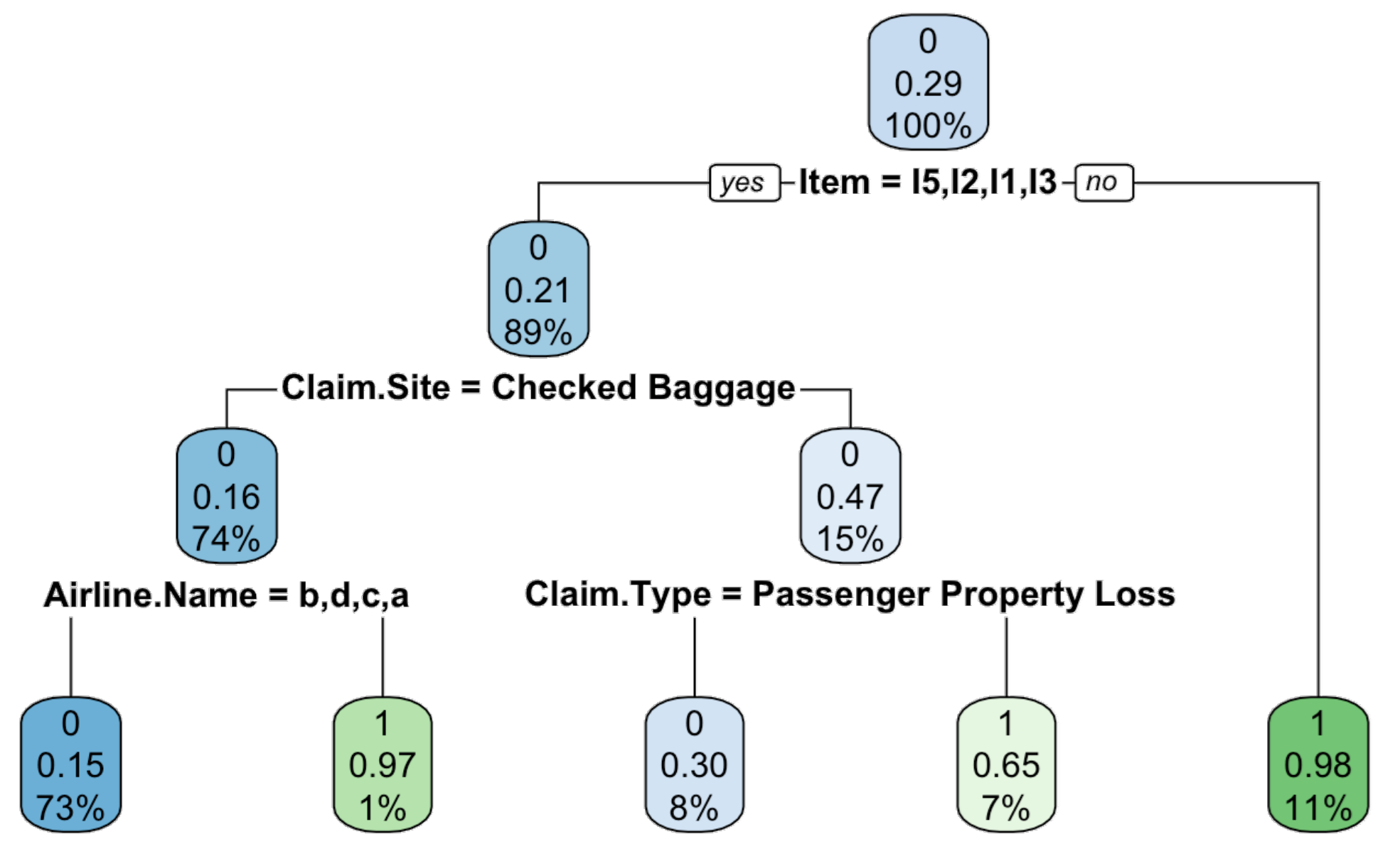

Figure 4: The decision tree

This tree can be interpreted in this way:

- If Item is in group I4, it is predicted the claim will be denied.
- If Claim.Site is Checked Baggage and Airline.Name is in group e, then it predicts the claim will be denied, if Airline.Name is in not in group e, then the claim will be approved.
- If Item is not in group 4, Claim.Site is not Checked Baggage, Claim.Type is Passegner Property Loss, then it predicts the claim will be denied; if Claim.Type is not Passegner Property Loss, then it is predicted the claim will be approved.

The variables that are used in making splits in this model are:

- Item
- Claim.Site
- Airline.Name
- Claim.Type

Boosted Forest

The variable importance is scaled to 100, which measures the relative importance of each variable. The more important variables usually appears earlier or more frequently in the tree splitting. It is clear the variables Item=4, Airline.Name=e, Claim.Site and Claim.Amount are more important than other variables. The Claim.Amount is important in this list but not shown in the tree in Task 2. This may be because the boosting method allows the variables that are overshadowed in the single tree to be fit to the errors made by other variables.The AUC of the boosted tree for the test data is 0.8694, which is promising. The variable importance plot is provided below:

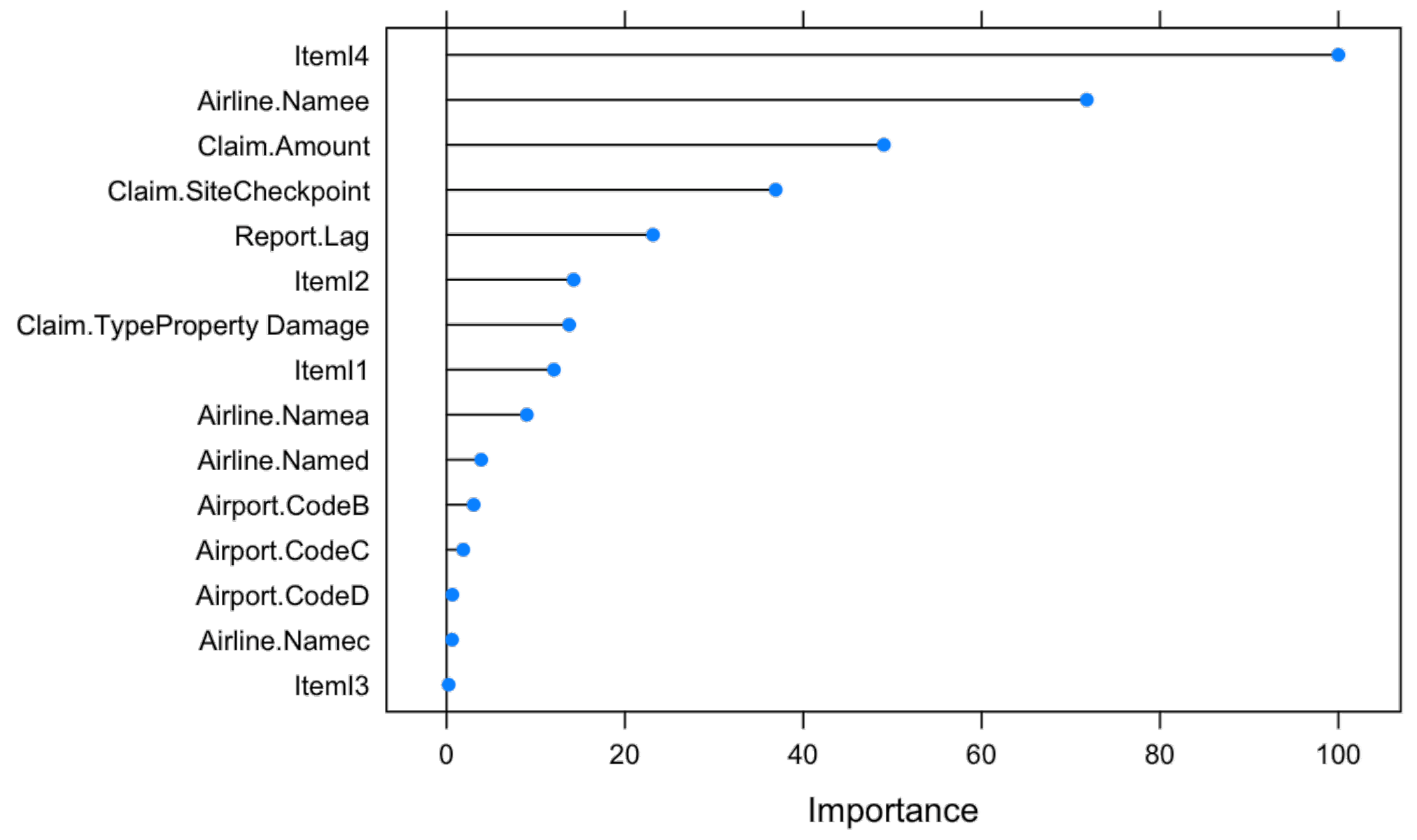

Figure 5: The boosted forest.The AUC of the boosted tree for the test data is 0.8694, which is promising.

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.

Here is the GLM coefficients using LASSO. We can see 4 features are removed.

(Intercept)	.
Report.Lag	.
Airport.CodeE	.
Airport.CodeC	0.15701822
Airport.CodeB	-0.10659870
Airport.CodeD	-0.02026289
Airline.Named	.
Airline.Namee	2.14630608
Airline.Namec	.
Airline.Namea	-0.05910761
Claim.TypeProperty Damage	0.10667057
Claim.SiteCheckpoint	0.58183343
ItemI2	-0.95894397
ItemI1	0.27954786
ItemI4	2.38364305
ItemI3	0.10974533
Claim.Amount_cutmediumAm	-0.15292703
Claim.Amount_cuthighAm	-0.19342332
Claim.TypeProperty Damage:Claim.SiteCheckpoint	0.98457488

Cutoff	Profit
0.20	25480
0.23	32195 (Best)
0.24	31480
0.25	30810
0.26	29765
0.30	29810
0.50	26850
0.60	14290

Table1: Cutoff values

The cutoff value is the probability threshold above which the prediction is classified as Approve, below which it's classified as Deny. Since the selection of cutoff values changes the classification results, I run the GLM at different cutoff to obtain the best cutoff that achieve 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.

Confusion Matrix

Below is the confusion matrix at cutoff=0.23.

	0	1
0	4329	632
1	860	1570

Table2: Confusion matrix

The column headers are the real classification, the row headers are the predictions. The first row first column is the number of True Negative (TN) predictions. The first row second column is the number of False Negative (FN). The second row first column is False Positive (FP). The second row second column is True Positive (TP). We will market to the positive predictions, and not market to the negative predictions. For positive prediction, if it is indeed positive, i.e. it is TP, we get a profit of 50. If it is indeed a negative, i.e. it is FP, we get a loss of 25. For the negative predictions, we don't market to them and get a loss of 5. Therefore, the total profit is 50*TP-25*FP-5*(TN+FN). The number of TP, FP, TN, FN changes if the cutoff changes. We tried different cutoff values and found out when cutoff=0.25, the total profit can be maximized.

A Model Demo for Marketing

The following table contains 1 base case and 7 other variation cases from the base case. To do a prediction, we need 7 pieces of information in the input, including Report.Lag, Aiport.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 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.

Report.Lag	Airport.Code	Airline.Name	Claim.Type	Claim.Site	Item	Claim.Amount	Probability	Prediction
17	MDW	Delta Air Lines	Passenger Property Loss	Checked Baggage	Clothing – Shoes; belts; accessories; etc.	65	0.1741	Deny
1	MDW	Delta Air Lines	Passenger Property Loss	Checked Baggage	Clothing – Shoes; belts; accessories; etc.	65	0.1741	Deny
17	**ABQ**	Delta Air Lines	Passenger Property Loss	Checked Baggage	Clothing – Shoes; belts; accessories; etc.	65	0.1741	Deny
17	MDW	**Air Canada**	Passenger Property Loss	Checked Baggage	Clothing – Shoes; belts; accessories; etc.	65	0.1741	Deny
17	MDW	Delta Air Lines	**Property Damage**	Checked Baggage	Clothing – Shoes; belts; accessories; etc.	65	0.1899	Deny
17	MDW	Delta Air Lines	Passenger Property Loss	**Checkpoint**	Clothing – Shoes; belts; accessories; etc.	65	0.2738	Approve
17	MDW	Delta Air Lines	Passenger Property Loss	Checked Baggage	**Locks**	65	0.0747	Deny
17	MDW	Delta Air Lines	Passenger Property Loss	Checked Baggage	Clothing – Shoes; belts; accessories; etc.	**3000**	0.1480	Deny

Table2: A spreadsheet demo of the developed model

In the second row of the above table, the case has one variation from the base case, which is changing from the Report.Lag from 17 to 1. The classification probability actually doesn't change. This is because the coefficients is 0 for variable Report.Lag in the GLM with LASSO features selection in Task 8. So the value of Report.Lag doesn't matter. As another example, the coefficients of Claim.Site=Checkpoint is 0.5818. Therefore, when the value of Claim.Site changes from "Checked Baggage" to "Checkpoint", the classification probability is increased to 0.2738.